Scuola di Dottorato
Università degli Studi di Milano-Bicocca

Department of

Economics, Management, and Statistics

Ph.D. program: **Economics and Statistics** Cycle: **XXXV°**
Curriculum: **Statistics**

Statistical Modeling and Temporal Clustering of Multivariate Time-Series with Applications to Financial Data

Surname: **Cortese**
Name: **Federico Pasquale**
Registration number: **854287**

Supervisor: Prof. **Fulvia Pennoni**
Co-Supervisors: Prof. **Francesco Bartolucci, Petter Kolm, Erik Lindström**

Academic Year: 2022-2023

[...]
Perché l'anima è in te, sei tu, ma tu
sei mia madre e il tuo amore è la mia schiavitù:
ho passato l'infanzia schiavo di questo senso
alto, irrimediabile, di un impegno immenso.
[...]
Ti supplico, ah, ti supplico: non voler morire.
Sono qui, solo, con te, in un futuro aprile...

"Supplica a mia madre" - P. P. Pasolini, 1962

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline and main contribution

Multivariate time-series models help to understand and predict how multiple variables change over time, particularly when they are interconnected and influence each other's dynamic. Temporal clustering helps in analyzing such data by identifying sets of time-series that exhibit similar patterns over time. It can be advantageous in detecting trends and anomalies, as well as making predictions for future observations. Precise modeling and clustering of time-series data can greatly enhance forecasting accuracy and facilitate effective anomaly detection. By contrast, improper specification of the statistical model can lead to incorrect grouping and loss of important information. In fact, inaccurate models can result in flawed interpretation of results, particularly in fields like finance, where misinterpretation can have significant repercussions (see Liao (2005) for a survey on clustering techniques for time-series data).

The modeling and clustering of financial time-series requires particular attention given their peculiarities, such as volatility clustering, heavy tails, non-linearity, and non-normality (Granger and Ding, 1994; Cont, 2001; Bulla and Bulla, 2006; Nystrup et al., 2015). Traditional approaches, such as the mean-variance analysis of Markowitz and Todd (2000) or the random walks, assume a normal distribution for the data, which fails to correctly estimate the probability of tail events (Fischer et al., 2009). Furthermore, classical clustering models may not be able to capture these trends, potentially resulting in incorrect estimates. Hence, advanced statistical and machine learning models are required for addressing these challenges.

High dimensionality usually characterizes this type of data (Liao, 2005; Aghabozorgi et al., 2015). A major problem is the curse of dimensionality, wherein an increasing number of variables leads to an exponential rise in the amount of data necessary to maintain the same level of accuracy. Moreover, high-dimensional data can lead to overfitting, which occurs when the model used for the analysis is too complex and captures noise in the data

rather than the underlying patterns. To address these issues, researchers have proposed various techniques, such as feature selection, dimensionality reduction, and regularization. The first involves selecting a subset of the most relevant variables for clustering analysis. Dimensionality reduction techniques, such as principal component analysis, can reduce the number of variables by projecting them onto a lower-dimensional space, while preserving the underlying structure of the data. Regularization techniques, such as Lasso (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970), can help to avoid overfitting by penalizing overly complex models.

Financial time-series present an additional complexity given their non-stationary nature, implying that the characteristics of such data vary over time. Non-stationarity can arise due to a variety of factors, such as changes in market sentiment, regulatory and economic environments, or unexpected events (Hamilton, 1989). For example, during a period of high market volatility or a financial crisis, the statistical properties of financial data can change abruptly, leading to different clustering results.

State-space models (Hamilton, 1994) have been recognized as a particularly suitable choice for effectively modeling non-stationary time-series data. These models are designed to represent the underlying dynamics of a system by modeling a set of unobserved states that evolve over time. At the heart of a state-space model is a set of two equations: the state equation, which describes how the underlying state of the system evolves over time, and the observation equation, which describes how the observed data are generated from the underlying state. The state equation is often formulated as a linear or non-linear dynamical system, while the observation equation is typically a function of the state and some noise process. State-space models have a number of advantages over more classical time-series models, including the ability to handle complex non-linear dynamics and non-Gaussian noise, or to incorporate prior knowledge about the system being modeled (Bartolucci and De Luca, 2003). They can accommodate temporal clustering by grouping the unobserved states, for instance using the Viterbi (1967) algorithm; see Pennoni and Bal-Domńska (2022) for an illustration of this approach with reference to longitudinal data and Pennoni et al. (2021) for time-series data. These models, with hidden Markov models (Zucchini et al., 2017; Bartolucci et al., 2022) being a popular example within this category, are particularly useful for modeling financial time-series because they are flexible enough and can help to analyze complex data structures. One advantage is their ability to separate the underlying state of the system from the noisy observations. In the context of financial time-series, states can be useful to infer market regimes, while the realized observations represent noisy measurements of these states.

This thesis discusses the modeling and clustering of multivariate financial time-series considering two specific model types that belong to the broader class of state-space models: *regime switching copula models* (Rodriguez, 2007; Nasri and Rémillard, 2019) and *sparse statistical jump models* (Bemporad et al., 2018; Nystrup et al., 2020, 2021). These models use temporal clustering techniques to effectively capture the dynamics of financial time-series data and overcome the problem of non-stationarity. Regime switching copula models are specifically designed to model the heavy tails of the joint distribution of financial returns, while sparse statistical jump models are ideal for problems involving a large number of variables.

Regime switching copula models involve two stochastic processes: one corresponds to the observed series, usually returns, while the other describes the evolution of an hidden sequence of states over time, which is modeled through a homogeneous Markov chain of first order. The joint distribution of returns is modeled as a copula function, which changes according to the underlying states. To address certain characteristics of this type of data and to model heavy-tails and non-linear dependencies, we propose a regime switching Student-t copula model. This model specifies the copula distribution as a Student-t copula (Demarta and McNeil, 2005), enhancing the detection of shifts between bullish and bearish market states.

Sparse statistical jump models allow for occasional jumps in asset returns, while maintaining an infrequent occurrence of these switches. They have the ability to choose the most relevant features that can effectively explain a complex system out of a large dataset. These models provide various advantages. Firstly, they can carry out parameter estimation, state-sequence decoding, and feature selection simultaneously. Secondly, they are highly robust towards incorrect model specifications and initialization, produces satisfactory results even with limited data samples, and are more efficient when dealing with high-dimensional feature vectors. Lastly, the estimated models are straightforward to interpret through their state-conditional dynamics and the weight assigned to each feature.

In the following sections, we briefly introduce each proposal, highlighting its main advantages. Additionally, we provide a brief overview of our findings and contributions, which are thoroughly discussed in subsequent Chapters. Specifically, Chapter 2 introduces a novel maximum likelihood estimation approach for regime switching Student-t copula models. We validate this proposal through a simulation study and apply it to jointly analyze the returns of five cryptocurrencies, with the goal of tracking their dynamics. In Chapter 3, we use sparse statistical jump models to identify the main drivers of the cryptocurrency market from a large set of features, including cryptocurrency returns, volatility, sentiment, and traditional financial assets. In Chapter 4, we implement a model selection method employing generalized information criteria within the framework of sparse statistical jump

models. We evaluate the effectiveness of this approach through an extensive simulation study, demonstrating its capacity to select the appropriate model. We further demonstrate the applicability of our approach in the context of the equity market, examining the dynamics and discovering the principal drivers of MSCI developed and emerging market indexes.

1.1.1 Regime switching Student-t copula model

Accurate modeling of the tails of the joint distribution of log-returns is crucial for effectively capturing the occurrence of extreme events. These events, which are more frequent than expected under a normal distribution, can result in significant losses for investors who fail to appropriately account for them (Christoffersen, 2011).

Copula models are particularly useful for modeling the tails of the joint distribution. They allow us to separate the marginal distributions from the underlying dependence structure, enabling a detailed analysis of individual variables dynamics while accurately capturing their interdependence (Sklar, 1959; Joe and Xu, 1996). Moreover, copulas offer great flexibility in modeling dependencies, as they assume different analytical forms that can capture many types of dependence.

The proposed regime switching Student-t copula model allows for the analysis of dynamic dependence structures between two or more random variables across different market conditions over time (Jondeau and Rockinger, 2006; Rodriguez, 2007). This variability is captured through a homogeneous Markov chain of first order, while the joint distribution is characterized through a Student-t copula.

The use of a Student-t copula distribution is advantageous in financial time-series modeling, as evidenced by various research papers, such as Breymann et al. (2003), Fischer et al. (2009), and Huang et al. (2009). By selecting the Student-t specification, we model dependencies through the number of degrees of freedom and the dependence parameter matrices. The number of degrees of freedom governs the thickness of the tails of the joint distribution, while the dependence parameter matrices allow to estimate the correlation structure.

Maximum likelihood estimation, especially through the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), is a widely used approach within regime switching copula models. However, the adoption of this algorithm can be challenging when more than two variables are considered. As the number of variables increases, the complexity of the model grows, and the estimation process becomes computationally demanding. To account for this issue, we propose a novel EM algorithm which is an extension of the approach developed by Trede (2020) for a simple Student-t copula model. We address the challenges of estimating the multidimensional Student-t copula parameters by introducing an iterative procedure

that efficiently estimates the matrix of dependence parameters and the number of degrees of freedom for a multivariate regime switching Student-t copula model. The proposed method is shown to be simple, computationally feasible, and fast, making it suitable for the analysis of multivariate financial data. We also show the good finite sample properties of the proposed estimator through simulations and we establish its computational efficiency, and its ability to yield better forecasting results as compared to traditional models.

We illustrate the feasibility of the proposal by modeling the log-returns of the five cryptocurrencies Bitcoin, Ethereum, Ripple, Litecoin and Bitcoin Cash over a five years period. Our analysis reveals that a two-state regime switching Student-t copula model accurately describes switches between bull and bear market regimes by examining the strength of correlations and the thickness of the tails of log-returns joint distribution.

1.1.2 Sparse statistical jump model

High dimensionality of financial data can make it challenging to estimate statistical models that accurately capture the association between variables. The sparse statistical jump model of Nystrup et al. (2021) addresses dataset sparsity, a situation where most variables have minimal influence on the model outcome, while a few key variables play a significant role. It is based on the so-called *statistical jump model*, introduced in Bemporad et al. (2018). Within their framework, they construct complex models by combining simpler ones, and the transition between these models is determined by an underlying latent process. The latter refers to the state of a given market, which can be understood as the current condition or configuration of the market's variables and dynamics. The model incorporates a specific hyperparameter that governs the level of persistence within a state, allowing for precise calibration of the model responsiveness to market changes. The sparse statistical jump model enhances the jump model framework by considering that some variables in the dataset are irrelevant for explaining the system dynamics. This assumption is integrated into the modeling process to improve accuracy by reducing dimensionality, and it is regulated by a specific hyperparameter controlling the level of sparsity.

Sparse statistical jump models have several advantages as they allow to jointly perform parameter estimation, feature selection, and state-sequence decoding. Compared to other competitive models, they are more efficient when working with high-dimensional feature vectors, more robust to misspecification errors and perform good even with small samples. Additionally, they are easy to interpret, based on their state-conditional dynamics.

The application we present focuses on the cryptocurrency market, with the added challenge of working with a high-dimensional dataset. Specifically, we consider over four hundred variables encompassing cryptocurrency returns, volatility and network activity, together with market sentiment, debt, equity, forex, and commodity market-related features.

By using the sparse statistical jump model, we find that a three-state model provides the best fit for describing changes in market dynamics. These states correspond to bull, neutral, and bear market conditions, each with clear and intuitive interpretations. Furthermore, we select the most relevant features for our analysis, including momentum, sentiment towards the crypto market, and trade activity.

1.1.3 Generalized information criteria for sparse statistical jump models

The selection of hyperparameters that control the number of states, sparsity, and persistence is a crucial challenge in the sparse statistical jump model framework. The number of states determines the complexity of the model and must be chosen carefully to avoid overfitting or underfitting. Similarly, the degree of sparsity needs to be precisely controlled to avoid excessive or inadequate dimensionality reduction. Additionally, regulating the states persistence is essential to capture the temporal dependencies of the features. Selecting these hyperparameters is often challenging, and a careful balance between model fit and complexity must be achieved to optimize the performance.

Information criteria are widely used in the context of model selection for hidden Markov models (Bacci et al., 2014). Their goal is to identify the optimal number of states and the appropriate model complexity that accurately captures the underlying data-generating process. Information criteria, such as *Akaike's information criterion* (Akaike, 1974) and *Bayesian information criterion* (Schwarz, 1978), offer a quantitative measure to compare models by balancing their goodness-of-fit and complexity. They provide a parsimonious and objective approach to identify the most suitable model by selecting the one with the lowest information criterion value.

The *generalized information criteria* (GIC) framework, introduced by Fan and Tang (2013), offers valuable guidance for model selection, particularly in high-dimensional scenarios. In fact, this framework extends conventional information criteria by incorporating a penalty term based on the number of variables included in the model, providing a flexible approach for model selection in high-dimensional scenarios.

We adapt the GIC framework to the class of sparse statistical jump models. Our approach involves deriving expressions for the model fit and complexity to construct suitable information criteria for model selection. To assess the performance of our proposal, we conduct two simulation studies. Our findings demonstrate that the proposed information criteria, tailored for sparse statistical jump models, identify the correct hyperparameter values. Specifically, our extended criteria correctly account for sparsity, persistence and provide an accurate estimate of the number of hidden states. These findings have important implications for researchers working with high-dimensional datasets, where model selection is a critical step in accurately representing the underlying data-generating process.

Thereafter, we conduct an empirical study wherein we apply a sparse statistical jump model with hyperparameters determined by our GIC. This model is applied to a large set of features associated with the global equity market. Our focus lies on MSCI indexes, encompassing developed and emerging markets, along with their cross-correlations and correlation with the VIX index. Leveraging the extended GIC, we detect the primary driver of equity markets, which predominantly consist of volatility-related features. Additionally, we trace its temporal dynamics, characterized by bull, neutral, and bear market phases.

1.2 Implications for real-world applications

Regime switching Student-t copulas and sparse statistical jump models present several benefits and implications for portfolio managers. They offer a powerful framework for understanding financial markets dynamics, aiding portfolio allocation strategies, managing tail risks, and uncovering different correlation patterns.

Enhancing portfolio allocations. These models capture the non-linear and time-varying nature of financial markets. By identifying different regimes characterized by distinct statistical properties, they allow for more accurate risk assessment and improved portfolio allocations. Studies like Hamilton (1989) and Gray (1996) underline the advantages of incorporating regime switching dynamics in portfolio optimization. Allocation strategies dynamically adjust asset distribution in response to market conditions, optimizing performance by aligning with probabilities of being in different market regimes.

Interpreting regimes. The interpretation of regimes is crucial. Different regimes could represent market bull/bear phases, high/low volatility periods, or various economic cycles. For instance, in a two-regime model, one regime might signify periods of economic expansion and low volatility, while the other denotes recessions and high volatility. Managers can then allocate assets accordingly, balancing risk and return profiles across different regimes (Ang and Chen, 2002).

Correlation patterns and implications. The proposed models uncover diverse correlation patterns across different market regimes. For portfolio managers, this means acknowledging that asset correlations can vary significantly during different market conditions. This insight can help diversify portfolios more effectively across regimes, reducing overall portfolio risk. Understanding these patterns, as highlighted by Garcia and Perron (1996) and Dahlquist and Gray (2000), enables managers to construct more robust and resilient portfolios towards changes in market conditions.

Enhancing tail risk management. Multivariate regime switching Student-t copula models are particularly useful in managing tail risks. In fact, these models capture extreme events and dependencies among asset returns during different regimes, allowing for a better understanding of tail risk dynamics. Tail risks, often underestimated by traditional models, can be more accurately assessed and managed by incorporating the multivariate nature of dependencies and tail behavior across regimes (Glosten et al., 1993).

1.3 Summary of the proposals

This thesis is the result of an extensive research aimed at advancing the field of modeling and clustering of multivariate financial time-series data. It is organized into three main Chapters, each contributing novel methods and insights to enhance the understanding of financial time-series analysis.

In Chapter 2, we present regime switching Student-t copula models and introduce a novel maximum likelihood estimation method for parameter estimation. We first provide a comprehensive introduction to regime switching Student-t copula models, showing their ability to capture extreme events and tail dependencies in multivariate financial time-series data. We discuss the challenges of estimating the parameters of these models, emphasizing the complexity introduced by the copula dependence structure. We then show our novel estimation method, emphasizing its ability to overcome computational challenges in estimating copula models with more than two variables. This improvement enhances both the efficiency and accuracy of parameter estimation. We demonstrate the practicality of our proposed method by applying it to cryptocurrency market data. This real-world application shows the effectiveness of our approach in capturing regime shifts and modeling the joint behavior of cryptocurrency returns.

Chapter 3 focuses on sparse statistical jump models and their application to the cryptocurrency market. We introduce sparse statistical jump models, designed to address the challenges posed by high-dimensional financial data. These models efficiently perform parameter estimation, feature selection, and state-sequence decoding simultaneously. We explore the unique characteristics of the cryptocurrency market considering a large set of candidate features related to cryptocurrency prices, cross-correlations, correlations with traditional financial assets, market activity and market sentiment. Our research highlights the importance of feature selection and the impact of selected features in modeling market conditions. Through an empirical analysis, we show that a three-state model effectively captures changes in cryptocurrency market conditions, characterizing each state as bull, neutral, and bear market regimes. We also show that momentum, trade activity, and

sentiment are the main drivers of this market, confirming prior research findings.

In Chapter 4, we introduce the concept of generalized information criteria and adapt its application to sparse statistical jump models, thus developing information criteria tailored for this model category. We emphasize the importance of model selection and illustrate how our criteria strike a balance between model fit and complexity. We conduct extensive simulation studies aimed at evaluating the performance of our proposal. The results demonstrate that the proposed approach identifies the correct number of latent states and improves model selection by effectively choosing the optimal values for the hyperparameters governing state transitions and dataset sparsity. Finally, providing an empirical application, we infer the key features that drive the return dynamics of the world equity market, which mainly consist of volatility-related features. We also find that a three-state model best describes the dynamics of MSCI developed and emerging markets indexes.

Bibliography

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–A decade review. *Information Systems*, 53:16–38.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Ang, A. and Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63:443–494.

Bacci, S., Pandolfi, S., and Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, 8:125–145.

Bartolucci, F. and De Luca, G. (2003). Likelihood-based inference for asymmetric stochastic volatility models. *Computational Statistics & Data Analysis*, 42:445–449.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and Its Application*, 9:425–452.

Bemporad, A., Breschi, V., Piga, D., and Boyd, S. P. (2018). Fitting jump models. *Automatica*, 96:11–21.

Breymann, W., Dias, A., and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3:1–14.

Bulla, J. and Bulla, I. (2006). Stylized facts of financial time-series and hidden semi-Markov models. *Computational Statistics & Data Analysis*, 51:2192–2209.

Christoffersen, P. (2011). *Elements of financial risk management*. Academic Press, San Diego, CA.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1:223–236.

Dahlquist, M. and Gray, S. F. (2000). Regime-switching and interest rates in the European monetary system. *Journal of International Economics*, 50:399–419.

Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73:111–129.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–22.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:531–552.

Fischer, M., Köck, C., Schlüter, S., and Weigert, F. (2009). An empirical analysis of multivariate copula models. *Quantitative Finance*, 9:839–854.

Garcia, R. and Perron, P. (1996). An analysis of the real interest rate under regime shifts. *The Review of Economics and Statistics*, 78:111–125.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48:1779–1801.

Granger, C. W. and Ding, Z. (1994). Stylized facts on the temporal and distributional properties of daily data from speculative markets. *UCSD Department of Economics Discussion Paper*, pages 94–19.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42:27–62.

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

Hamilton, J. D. (1994). State-space models. *Handbook of Eonometrics*, 4:3039–3080.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huang, J. J., Lee, K. J., Liang, H., and Lin, W. F. (2009). Estimating value at risk of portfolio by conditional copula-GARCH method. *Insurance: Mathematics and Economics*, 45:315–324.

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. *University of British Columbia, Department of Statistics, Technical Report*, 166.

Jondeau, E. and Rockinger, M. (2006). The copula-GARCH model of conditional dependencies: An international stock market application. *Journal of International Money and Finance*, 25:827–853.

Liao, T. W. (2005). Clustering of time-series data–A survey. *Pattern Recognition*, 38:1857–1874.

Markowitz, H. M. and Todd, G. P. (2000). *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons.

Nasri, B. R. and Rémillard, B. N. (2019). Copula-based dynamic models for multivariate time-series. *Journal of Multivariate Analysis*, 172:107–121.

Nystrup, P., Kolm, P. N., and Lindström, E. (2021). Feature selection in jump models. *Expert Systems with Applications*, 184:115558.

Nystrup, P., Lindström, E., and Madsen, H. (2020). Learning hidden Markov models with persistent states by penalizing jumps. *Expert Systems with Applications*, 150:113307.

Nystrup, P., Madsen, H., and Lindström, E. (2015). Stylised facts of financial time-series and hidden Markov models in continuous time. *Quantitative Finance*, 15:1531–1541.

Pennoni, F. and Bal-Domńska, B. (2022). NEETs and youth unemployment: A longitudinal comparison across European countries. *Social Indicator Research*, 162:739–761.

Pennoni, F., Bartolucci, F., Forte, G., and Ametrano, F. (2021). Exploring the dependencies among main cryptocurrency log-returns: A hidden Markov model. *Economic Notes*, 51:e12193.

Rodriguez, J. C. (2007). Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, 14:401–423.

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

Sklar, A. (1959). Fonctions de repartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288.

Trede, M. (2020). Maximum likelihood estimation of high-dimensional Student-t copulas. *Statistics & Probability Letters*, 159:108678.

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

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R*. CRC press, Boca Raton, FL.

Chapter 2

Maximum likelihood estimation of multivariate regime switching Student-t copula models[1]

2.1 Introduction

Financial analysis and risk management research shows that the dependence structure of financial time-series changes during crises, with interdependence among assets increasing compared to stable periods (Das and Uppal, 2004; Patton, 2004). This phenomenon, known as *asymmetric dependence* (Ang and Bekaert, 2002; Ang and Chen, 2002; Longin and Solnik, 2002), is particularly relevant in cryptocurrency markets due to their vulnerability to changes in economic developments and news (Garcia and Ghysels, 1998; Kristoufek, 2013; Telli and Chen, 2020). For this reason, it is important to consider suitable specifications for the joint distribution of log-returns to capture possible sudden changes in market dynamics.

Extensive research supports the existence of regime switches in cryptocurrency returns, volatilities, and cross-correlation structure; see Ardia et al. (2019), Shen et al. (2020) and Cremaschini et al. (2023). Hamilton (1989) first argued that the switching dynamics of financial returns may be easily modeled through a Markovian process. In particular, regime switching (RS) copula models, also known as Markov-switching copulas (Jondeau and Rockinger, 2006; Rodriguez, 2007; Okimoto, 2008; Chollete et al., 2009), accurately describe persistent correlation dynamics (Ang and Timmermann, 2012) in log-returns by modeling the joint distribution as a copula function that changes according to latent states. These models involve two stochastic processes: the first corresponds to the observed series, and the other to an underlying (latent) process describing the evolution of the hidden states over time. RS copula models are employed for exchange rates data, for the analysis of the correlation between S&P 500 and NASDAQ indexes, for the study of gold-oil dependence structure, and to describe momentum shifts in football matches (Stöber and Czado, 2014;

[1] This Chapter has been submitted for publication in the International Statistical Review and is currently undergoing minor revisions. Cortese F., Pennoni F., Bartolucci F. Maximum likelihood estimation of multivariate regime switching Student-t copula models.

Härdle et al., 2015; Nasri and Rémillard, 2019; Tiwari et al., 2020; Ötting et al., 2021).

In the present Chapter, we propose a new model, named RS Student-t copula (RSStC) model, tailored to account for stylized facts of financial returns, such as heavy-tailed distributions and non-linear dependencies. The model is based on a Student-t copula (Demarta and McNeil, 2005) which is parametrized by the number of degrees of freedom and the matrix of dependence parameters. Student-t copula is generally preferred to Gaussian copula for financial time-series because it allows the modeling of tail dependence and kurtosis (Breymann et al., 2003; Fischer et al., 2009; Huang et al., 2009). Compared to Archimedean copulas (Genest et al., 2011), which rely on a single dependence parameter for all variables, the proposed RSStC formulation offers superior accuracy in modeling the correlation structure. Additionally, while Vine copulas (Joe and Kurowicka, 2011; Czado and Nagler, 2022) are quite flexible to model different distributions, they have a more complex analytical form.

Maximum likelihood estimation of the RS copula parameters is typically performed through the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). However, even in the case of a simple multidimensional Student-t copula model, Hernández et al. (2014) show that maximum likelihood estimation can be highly computationally inefficient. We provide a new approximation method for the EM algorithm tailored for estimating RSStC models. The proposal consists in an iterative procedure for estimating the matrix of dependence parameters and the number of degrees of freedom of the multivariate RSStC model. Following the approach of Trede (2020), developed for estimating a simple Student-t copula model, we maximize the log-likelihood function corresponding to the Student-t copula density in two steps. At the first step, we estimate the matrix of dependence parameters for a fixed number of degrees of freedom through Lagrange multipliers relying on a closed form solution; then, we numerically optimize the log-likelihood with respect to the number of degrees of freedom, keeping fixed the estimated matrix of dependence parameters. This procedure is simple, computationally feasible, and fast, even for long series with many assets. To evaluate the proposal, we rely on a simulation study assessing the good finite sample properties of the estimates and the computational efficiency of the procedure. An important feature of the proposal is that it can account for persistence in market regimes. This is an important aspect since, as suggested in Nystrup et al. (2020), when the state sequence contains several jumps, the RS model tends to a finite mixture model (McLachlan and Peel, 2000).

We apply the proposed approach to analyze log-returns of the five cryptocurrencies, Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Bitcoin Cash (BTC), for five years, from 17 September 2017 to 02 October 2022. At least to our knowledge, these data have never been analyzed with RS copula models; for a recent review of the methods proposed in the literature for the analysis of multiple cryptoassets, see, among others, Koki

et al. (2022). We select the optimal number of latent states relying on the Integrated Completed Likelihood (ICL) criteria (Biernacki et al., 2000) and, following Pennoni et al. (2021), we predict the latent regimes considering global decoding (Viterbi, 1967; Juang and Rabiner, 1991). Additionally, we compare the forecasting performance of the proposed model with that of a more common hidden Markov model (HMM) (Zucchini et al., 2017) and a multivariate random walk (MRW) process considered as benchmark.

To summarize, we provide three main contributions to the existing literature. First, we propose a multivariate RSStC model, whereas most previous works focus only on the bivariate case. Second, we implement a novel and computationally efficient method for estimating the matrix of dependence parameters and the number of degrees of freedom. Third, we show the applicability of the proposal analyzing cryptocurrency log-returns in a novel way.

We implemented the code developed to carry out the estimation of the RSStC model and to perform simulations in the `C++` language, through the `R` (R Core Team, 2023) package **Rcpp** (Eddelbuettel and François, 2011). We make this code freely available at the following link: `https://github.com/FedericoCortese/RSstcopula/find/main`.

The remainder of the Chapter is organized as follows. In Section 2.2, we introduce the RSStC model. In Section 2.3, we show the proposed procedure for maximum likelihood estimation of the model parameters, the initialization strategy, and the convergence criterion chosen for the EM algorithm. In Section 2.4, we illustrate the simulation study aimed at assessing the validity of the proposed estimation procedure, and we comment on the results. In Section 2.5, we apply the proposal to analyze the daily log-returns of the five cryptocurrencies and present the results along with comparative analysis. In Section 2.6 we discuss the obtained results. In Appendix 2.A, we show additional details of the E- and M-steps of the EM algorithm, and we describe the model selection criterion. Appendix 2.B contains more details on the simulation results. In Appendix 2.C, we show an application in which the model is estimated using a semi-parametric approach.

2.2 Model formulation

We consider an r-dimensional copula function C, which is a multivariate cumulative distribution function on the hypercube $[0,1]^r$ with marginal uniform distributions in $[0,1]$. Estimating such a model through inferential procedures becomes challenging due to limitations in numerical optimization methods when dealing with high-dimensional parameter vectors. Additionally, the joint likelihood often involves multidimensional integrals, posing difficulties in numerical computations. To solve this problem, we rely on Sklar's theorem (Sklar, 1959), suggesting that it is possible to separately estimate each marginal cumulative distribution function and the copula function. The inference for margins approach of Joe

and Xu (1996) allows us to split the estimation into two steps: first, we fit the marginal distribution of each univariate time-series; second, we estimate the joint distribution of integral transforms of these series using a RS copula model.

In the following, for the sake of clarity we explicitly refer to the practical context of financial data. Let $\boldsymbol{y}_t = (y_{t1}, \dots, y_{tr})'$ denote the vector of log-returns of the r time-series at time $t = 1, \dots, T$. Following a parametric approach (Joe, 1997; Nasri and Rémillard, 2019), we assume a generalized error model (Du, 2016) for each of the r univariate time-series. This model postulates a cumulative distribution function denoted as $G_{\boldsymbol{\beta}j}$, and characterizes the integral transforms $z_{tj} = G_{\hat{\boldsymbol{\beta}}j}(y_{tj})$ as independent and identically distributed random variables with continuous distribution function F_j, $j = 1, \dots, r$. To eliminate the dependence of the estimated copula parameters on the marginal distributions, as suggested by Nasri and Rémillard (2019), we initially estimate the parameters $\boldsymbol{\beta}_j$ through a consistent estimator $\hat{\boldsymbol{\beta}}_j$, then we compute the uniform pseudo-observations z_{tj} and finally we calculate normalized ranks, denoted by $\hat{e}_{tj} = \text{rank}(z_{tj})/(T+1)$, for $t = 1, \dots, T$, $j = 1, \dots, r$. Following a semi-parametric approach, normalized ranks can be directly calculated from the observed log-returns, thus obtaining $\hat{e}_{tj} = \text{rank}(y_{tj})/(T+1)$. This is a valid alternative when there is no interest in estimating a parametric model for the marginal univariate time-series, which might be the case when the focus is only on the association between a set of random variables and not on their marginal distributions. We consider the first method in the application presented in Section 2.5 to analyze cryptocurrency log-returns, and we also offer the results with the semi-parametric approach in Appendix 2.C.

With a slight abuse of notation, let us denote with $\boldsymbol{y}_t = (y_{t1}, \dots, y_{tr})'$, $t = 1, \dots, T$, the r-dimensional vector of pseudo-observations following an RSStC model, and let u_t denote the latent variable assumed to follow a time homogeneous Markov process of first order with k latent states. We conceive the following assumptions:

- The latent process is characterized by a vector of initial probabilities $\boldsymbol{\lambda}$ with elements $\lambda_u = P(u_1 = u)$, $u = 1, \dots, k$, and a transition matrix denoted as $\boldsymbol{\Pi}$, with elements $\pi_{v|u} = P(u_t = v \mid u_{t-1} = u)$, $u, v = 1, \dots, k$.
- The vectors of pseudo-observations $\boldsymbol{y}_1, \dots, \boldsymbol{y}_T$ are conditionally independent given the latent regimes $u_1, \dots, u_T$, each with copula densities $c(\cdot; \boldsymbol{R}_{u_t}, \nu_{u_t})$, $t = 1, \dots, T$. $\boldsymbol{R}_u$ denotes the matrix of dependence parameters with entries $\rho_u^{(ij)}$, $i, j = 1, \dots, r$, $i \neq j$, each measuring the correlation between assets i and j, and ν_u is the number of degrees of freedom of the Student-t copula.

The joint density of the pseudo-observations is given by

$$f(\boldsymbol{y}_1, \dots, \boldsymbol{y}_T) = \sum_{u_1=1}^{k} \pi_{u_1} c(\boldsymbol{y}_1; \boldsymbol{R}_{u_1}, \nu_{u_1}) \cdot \dots \cdot \sum_{u_T=1}^{k} \pi_{u_T|u_{T-1}} c(\boldsymbol{y}_T; \boldsymbol{R}_{u_T}, \nu_{u_T}). \tag{2.1}$$

More specifically, following Joe (2014), $c(\boldsymbol{y}_t; \boldsymbol{R}_{u_t}, \nu_{u_t})$, $t = 1, \ldots, T$, is given by

$$c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u) = \frac{t_{r,\nu_u}(\boldsymbol{x}_t; \boldsymbol{R}_u)}{\prod_{j=1}^{r} t_{1,\nu_u}(x_{tj})}, \quad u = 1, \ldots, k,$$

where $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{rt})'$ is the vector with components $x_{tj} = T^{-1}_{1,\nu_u}(y_{tj})$, $j = 1, \ldots, r$, and T^{-1}_{1,ν_u} is the inverse cumulative distribution function of a one-dimensional Student-t random variable with ν_u degrees of freedom. The univariate and r-variate Student-t densities, denoted as t_{1,ν_u} and t_{r,ν_u}, are defined as

$$t_{1,\nu_u}(x_{tj}) = \frac{\Gamma((\nu_u+1)/2)}{\sqrt{\pi\nu_u}\,\Gamma(\nu_u/2)} \left(1 + \frac{x_{tj}^2}{\nu_u}\right)^{-(\nu_u+1)/2},$$

$$t_{r,\nu_u}(\boldsymbol{x}_t; \boldsymbol{R}_u) = \frac{\Gamma\left(\frac{\nu_u+r}{2}\right)}{\Gamma\left(\frac{\nu_u}{2}\right) \nu_u^{\frac{r}{2}} \pi^{\frac{r}{2}} |\boldsymbol{R}_u|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu_u} \boldsymbol{x}_t^T \boldsymbol{R}_u^{-1} \boldsymbol{x}_t\right)^{-\frac{\nu_u+r}{2}},$$

where $\Gamma(\cdot)$ is the gamma function.

To measure correlation between variables, we employ the Kendall's tau (Kendall, 1938), which offers advantages over the linear correlation coefficient, particularly in its ability to model non-linear dependencies. Starting from the dependence parameters $\rho_u^{(ij)}$ of the Student-t copula, it is easy to compute Kendall's tau through the following formula:

$$\tau_u^{(ij)} = \frac{2}{\pi} \arcsin \rho_u^{(ij)}, \quad i, j = 1, \ldots, r, \quad i \neq j. \tag{2.2}$$

2.3 Maximum likelihood estimation

Let $\ell(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ denote the log-likelihood of the proposed model, corresponding to the logarithm of (2.1), with $\boldsymbol{\theta}$ being the column vector of parameters including the non-redundant elements of $\boldsymbol{R}_u$, together with ν_u and λ_u, for $u = 1, \ldots, k$, and $\pi_{v|u}$, for $u, v = 1, \ldots, k$. Maximum likelihood estimation of the model parameters is performed through the EM algorithm; for the Student-t copula parameters, we use a two-step procedure where we first estimate the matrix of dependence parameters for a fixed number of degrees of freedom, through Lagrange multipliers, and then we estimate the number of degrees of freedom through numerical optimization of the complete log-likelihood, keeping fixed the previously estimated matrix. In the following section, we show the steps of the EM algorithm, details of which are provided in Appendix 2.A. In Section 2.3.2, we provide additional information on the initialization of the algorithm and its convergence.

2.3.1 Expectation-Maximization algorithm

The complete-data log-likelihood, denoted as $\ell^*(\boldsymbol{\theta} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T))$, is the log-likelihood computed assuming the knowledge of the hidden states $u_1, \ldots, u_t$, and expressed as

$$\begin{aligned}\ell^*(\boldsymbol{\theta} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T)) &= \sum_{t=1}^{T}\sum_{u=1}^{k} w_{tu} \log c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u) + \sum_{u=1}^{k} w_{1u} \log \lambda_u \\ &\quad + \sum_{t=2}^{T}\sum_{u=1}^{k}\sum_{v=1}^{k} z_{tuv} \log \pi_{v|u},\end{aligned} \tag{2.3}$$

being $w_{tu} = I(u_t = u)$ an indicator variable equal to 1 when the latent process is in state u at time t (0 otherwise), and $z_{tuv} = I(u_{t-1} = u, u_t = v)$ equal to 1 if the latent process switches from state u at time $t-1$ to state v a time t (0 otherwise). Note that this log-likelihood is the sum of three components that may be maximized separately.

Starting from some initial values for the parameters collected into the vector $\boldsymbol{\theta}^{(0)}$, the EM-algorithm maximizes $\ell(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ by alternating the following two steps until convergence:

- **E-step**. Compute the conditional expected value of $\ell^*(\boldsymbol{\theta} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T))$, given the values of the parameters at the previous iteration and the pseudo-observations. At this step, we rely on the posterior expected values of the previous indicator variables, denoted by $\hat{w}_{tu}$ and $\hat{z}_{tuv}$, whose formulas are provided in Appendix 2.A.

- **M-step**. Maximize the expected value of $\ell^*(\boldsymbol{\theta} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T))$ and update the model parameters. In particular, parameters λ_u and $\pi_{v|u}$ are updated by using the following explicit rules:

$$\lambda_u^{(m)} = \frac{\hat{w}_{1u}}{\sum_{v=1}^{k} \hat{w}_{1v}}, \quad u = 1, \ldots, k, \tag{2.4}$$

$$\pi_{v|u}^{(m)} = \frac{\sum_{t=2}^{T} \hat{z}_{tuv}}{\sum_{t=2}^{T} \hat{w}_{t-1u}}, \quad u, v = 1, \ldots, k. \tag{2.5}$$

  The updated values of the remaining parameters, that is, $\boldsymbol{R}_u^{(m)}$ and $\nu_u^{(m)}$, are obtained by solving the following optimization task

$$\max_{\boldsymbol{R}_u, \nu_u} \sum_{t=1}^{T} \hat{w}_{tu} \log c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u), \quad u = 1, \ldots, k. \tag{2.6}$$

A naive, direct numerical maximization of (2.6) may be performed. However, it results computationally inefficient, especially when the number of available assets is large. Following Trede (2020), we maximize (2.6) with respect to $\boldsymbol{R}_u$, $u = 1, \ldots, k$, given $\nu_u^{(m-1)}$ using Lagrange multipliers, to obtain $\boldsymbol{R}_u^{(m)}$, and then with respect to ν_u given $\boldsymbol{R}_u^{(m)}$, obtaining

$\nu_u^{(m)}$. In particular, we obtain an estimated approximation of the matrix of dependence parameters via the following rule:

$$\boldsymbol{R}_u^{(m)} = \boldsymbol{A}^{(m-1)} + \boldsymbol{R}_u^{(m-1)} \operatorname{diag}\left[\left(\boldsymbol{R}_u^{(m-1)} \circ \boldsymbol{R}_u^{(m-1)}\right)^{-1}\left(\mathbf{1} - \boldsymbol{a}^{(m-1)}\right)\right] \boldsymbol{R}_u^{(m-1)}, \tag{2.7}$$

with

$$\boldsymbol{A}^{(m-1)} = \frac{\nu_u^{(m-1)} + r}{\nu_u^{(m-1)} \sum_t \hat{w}_{tu}} \sum_{t=1}^{T} \hat{w}_{tu} \boldsymbol{x}_t \boldsymbol{x}_t' \left[1 + \frac{\boldsymbol{x}_t' \left(\boldsymbol{R}_u^{(m-1)}\right)^{-1} \boldsymbol{x}_t}{\nu_u^{(m-1)}}\right]^{-1},$$

where $\circ$ in (2.7) denotes the element-wise product, $\mathbf{1}$ is a vector of 1s, $\boldsymbol{a}^{(m-1)}$ denotes the vector of diagonal elements of $\boldsymbol{A}^{(m-1)}$, and $\boldsymbol{x}_t$ is the vector with components $T_{1,\nu_u}^{-1}(y_{tj})$, with $j = 1, \ldots, r$. In this way, we do not require numerical optimization methods for such an estimate, thus reducing the computational effort. See Appendix 2.A for additional details on the derivation of the above formulas.

It is computationally simple to numerically maximize Equation (2.6) with respect to ν_u once we set $\boldsymbol{R}_u = \boldsymbol{R}_u^{(m)}$, because the number of degrees of freedom ν_u of the Student-t copula is a scalar parameter. The estimates for ν_u, $u = 1, \ldots, k$, are obtained as

$$\nu_u^{(m)} = \underset{\nu_u}{\operatorname{argmax}} \sum_{t=1}^{T} \hat{w}_{tu} \log c(\boldsymbol{y}_t; \boldsymbol{R}_u^{(m)}, \nu_u), \quad u = 1, \ldots, k. \tag{2.8}$$

We employ a heuristic approach with specific bounds to ensure successful computation of the objective function for estimating the parameter ν_u. We set the lower bound at 2 for practical purposes, and the upper bound is chosen as 25 to prevent numerical instability of the algorithm at higher values. As also reported in Trede (2020), larger values of ν_u imply a significantly higher computational time needed to achieve convergence to the maximum of the log-likelihood function. Additionally, as the number of degrees of freedom increases, the Student-t copula gradually approximates the Gaussian copula, and it may result less effective in capturing extreme returns. In the simulation study of Section 2.4 and in the application of Section 2.5, we arrange these regimes in ascending order of determinants, from 1 to 0, corresponding to decreasing "general" correlation values.

Standard errors for the parameter estimates are computed by parametric bootstrap (Davison and Hinkley, 1997; Chernick, 2011). In particular, we make use of the stationary block bootstrap (Politis and Romano, 1994) to preserve time-series dependence of the data: it consists in resampling blocks of consecutive observations and the length of each block is distributed as a geometric random variable with average size proportional to $\mathcal{O}(T^{2/3})$.

2.3.2 Initialization and convergence of the algorithm

In the literature, there is currently no consensus on the most appropriate approach for initializing the values in $\boldsymbol{\theta}^{(0)}$ within the context of the EM algorithm. We follow the proposal in Bartolucci et al. (2013) (Chapter 3) and we use a deterministic rule as an initialization strategy for $\boldsymbol{R}_u$, $u = 1, \ldots, k$, such that initial values are defined on the basis of the descriptive statistics computed for the observed time-series. To determine the initial values for the dependence parameters, we begin by computing the matrix of sample Kendall's tau. Subsequently, we invert the formula in Equation (2.2) to obtain initial estimates of $\rho_u^{(ij)}$. Other possible choices are illustrated in Maruotti and Punzo (2021). The starting values for the initial probabilities λ_u are set equal to $1/k$, and those of the transition probabilities $\pi_{v|u}$ are set equal to $1/(\gamma + k)$ for $v \neq u$ and equal to $(h+1)/(\gamma + k)$ for $v = u$, where γ is a suitable constant (we use $\gamma = 0$ in our application). We note that a moderate initial value of ν_u is the best practical choice: it should not be too large or too small because we might encounter convergence issues. For this reason, based on a heuristic strategy, the number of degrees of freedom of the Student-t copula is initialized with 4.

Regarding algorithm convergence, we employ two common approaches: monitoring the distance between estimated parameter vectors at consecutive steps and tracking the increase in the log-likelihood function at each step. Specifically, the E- and M-steps iterate until either or both of the following conditions are met

$$\max_h \left| \theta_h^{(m+1)} - \theta_h^{(m)} \right| < \epsilon_1,$$

$$\left| \ell(\boldsymbol{\theta}^{(m+1)}) - \ell(\boldsymbol{\theta}^{(m)}) \right| < \epsilon_2,$$

being $\theta_h^{(m)}$ the h-th element of the vector $\boldsymbol{\theta}^{(m)}$ at the m-th iteration of the algorithm and $\epsilon_1, \epsilon_2 > 0$, suitable tolerance levels. In the simulation study presented in Section 2.4 and in the empirical analysis of Section 2.5, both tolerance levels are set at 10^{-8}.

2.4 Simulation study

We validate the proposed RSStC model through a simulation study, examining the properties of the estimators for dependence parameters, degrees of freedom, initial and transition probabilities. In particular, we present the simulation results for a 3-state RSStC model.

We conduct experiments on a Standard NC6 Promo virtual machine with 6 cores and 56 GB of memory. We acknowledge the University of Milano-Bicocca Data Science Lab (data-lab) for supporting this work by providing some computational resources. As mentioned in the introduction, we implement the `R` code for the EM algorithm through the package **Rcpp**,

which allows the user to easily integrate `C++` into the `R` environment. The code is available at the following link https://github.com/FedericoCortese/RSstcopula/find/main.

The estimation procedure is remarkably efficient, as it takes only around 50 seconds to estimate a 3-state model with data consisting of $1,500$ observations and 5 marginals. Moreover, the computational time increases linearly with both the dataset size (T) and the number of states (r).

2.4.1 Three state regime switching Student-t copula model

We generate data from a 3-state RSStC model drawing $B = 1,000$ samples of dimension $r = 5$, each with a total number of observations $T =$ 1,500. The model has equally probable initial probabilities $\boldsymbol{\lambda} = (1/3, 1/3, 1/3)'$ and transition probability matrix given by

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.700 & 0.200 & 0.100 \\ 0.300 & 0.600 & 0.100 \\ 0.100 & 0.100 & 0.800 \end{bmatrix}.$$

The dependence matrices are fixed at

$$\boldsymbol{R}_1 = \begin{bmatrix} 1.000 & - & - & - & - \\ 0.900 & 1.000 & - & - & - \\ 0.700 & 0.750 & 1.000 & - & - \\ 0.800 & 0.900 & 0.700 & 1.000 & - \\ 0.800 & 0.800 & 0.800 & 0.800 & 1.000 \end{bmatrix}, \quad \boldsymbol{R}_2 = \begin{bmatrix} 1.000 & - & - & - & - \\ 0.500 & 1.000 & - & - & - \\ 0.300 & 0.400 & 1.000 & - & - \\ 0.500 & 0.400 & 0.400 & 1.000 & - \\ 0.400 & 0.500 & 0.500 & 0.300 & 1.000 \end{bmatrix},$$

$$\boldsymbol{R}_3 = \begin{bmatrix} 1.000 & - & - & - & - \\ 0.100 & 1.000 & - & - & - \\ 0.150 & -0.100 & 1.000 & - & - \\ 0.050 & 0.100 & 0.050 & 1.000 & - \\ 0.050 & -0.050 & 0.100 & -0.010 & 1.000 \end{bmatrix},$$

with state-specific numbers of degrees of freedom equal to $\nu_1 = 3$, $\nu_2 = 6$ and $\nu_3 = 10$, respectively.

We evaluate the estimator in terms of the average bias and root mean squared error (RMSE) across $B = 1,000$ samples computed for the h-th parameter θ_h as

$$\text{Bias} = \text{E}\left(\tilde{\theta}_h - \theta_h\right),$$

$$\text{RMSE} = \sqrt{\text{E}\left(\tilde{\theta}_h - \theta_h\right)^2 + \text{Var}\left(\tilde{\theta}_h\right)},$$

where $\tilde{\theta}_h$ denotes the h-th component of the vector of parameters $\tilde{\boldsymbol{\theta}}$ estimated on the simulated sample, and θ_h the corresponding component of the vector $\boldsymbol{\theta}$ of true parameters. We also compute bootstrap percentiles confidence intervals (CI) at 95% level.

Table 2.1 shows results for the parameters of the hidden Markov process under different

scenarios. The bias is always low and RMSEs of the initial probabilities are around 0.5, while they are below 0.168 for the transition probabilities. CIs of the transition probabilities are narrower for higher probabilities.

Table 2.1: Simulation results for the 3-state RSStC model: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	λ_3	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|3}$	$\pi_{2\|1}$	$\pi_{2\|2}$	$\pi_{2\|3}$	$\pi_{3\|1}$	$\pi_{3\|2}$	$\pi_{3\|3}$
True	0.333	0.333	0.333	0.700	0.300	0.100	0.200	0.600	0.100	0.100	0.100	0.800
Bias	-0.014	0.079	-0.065	0.019	-0.012	-0.007	-0.027	0.085	-0.058	0.000	-0.025	0.026
RMSE	0.475	0.440	0.492	0.080	0.074	0.048	0.099	0.168	0.112	0.040	0.080	0.074
CI_L	0.000	0.000	0.000	0.461	0.100	0.024	0.167	0.225	0.019	0.027	0.015	0.603
CI_U	1.000	1.000	1.000	0.774	0.354	0.220	0.517	0.730	0.384	0.180	0.299	0.859

Table 2.2 reports results of the dependence parameters $\rho_u^{(ij)}$ and the number of degrees of freedom ν_u. The maximum absolute bias for the dependence parameters is 0.037, specifically for the pair of marginals $(2, 3)$ in the second state. RMSE values are all below 0.135. CIs for the first state are narrower, indicating more accurate estimates when the correlation is high, while CIs for the second state show higher uncertainty in the estimated parameters. As ν_u decreases, bias tends to decrease, resulting in improved estimation results when fitting distributions with fat tails.

We also examined a 2-state RSStC model in a separate simulation study, and the outcomes closely resemble those of the 3-state model. Furthermore, we varied the number of observations, T, and the number of assets, r. Our findings indicate that the proposed approach demonstrates good finite sample properties, as evidenced by a decline in RMSE with increasing T. Similarly, as r increases, the RMSE decreases for initial and transition probabilities, as well as for the number of degrees of freedom, while it shows an increment for dependence parameters. Comprehensive information on these simulation results can be found in Appendix 2.B.

2.5 Empirical study

2.5.1 Data

Data used for the application are multidimensional time-series of the daily log-returns considered at closing prices of BTC, ETH, XRP, LTC, and BCH, which are, in terms of market capitalization, the less manipulated and more liquid crypto assets. We acknowledge the Crypto Asset Lab, which is an independent academic lab established at the University

Table 2.2: Simulation results for the 3-state RSStC model: the true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.700	0.800	0.800	0.750	0.900	0.800	0.700	0.800	0.800	3.000
Bias	0.001	0.003	0.002	0.002	0.003	0.001	0.003	0.003	0.002	0.002	-0.031
RMSE	0.015	0.034	0.025	0.026	0.030	0.015	0.028	0.034	0.025	0.026	0.441
CI_L	0.869	0.628	0.750	0.746	0.688	0.867	0.742	0.635	0.747	0.745	2.285
CI_U	0.924	0.763	0.842	0.843	0.802	0.923	0.843	0.759	0.842	0.843	4.054
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.500	0.300	0.500	0.400	0.400	0.400	0.500	0.400	0.500	0.300	6.000
Bias	0.032	0.032	0.031	0.032	0.037	0.035	0.032	0.026	0.029	0.036	0.611
RMSE	0.110	0.117	0.110	0.119	0.124	0.128	0.115	0.109	0.110	0.135	1.844
CI_L	0.302	0.108	0.310	0.197	0.206	0.175	0.302	0.228	0.320	0.194	3.848
CI_U	0.759	0.581	0.741	0.660	0.691	0.719	0.760	0.656	0.736	0.621	10.395
State 3	$\rho_3^{(12)}$	$\rho_3^{(13)}$	$\rho_3^{(14)}$	$\rho_3^{(15)}$	$\rho_3^{(23)}$	$\rho_3^{(24)}$	$\rho_3^{(25)}$	$\rho_3^{(34)}$	$\rho_3^{(35)}$	$\rho_3^{(45)}$	ν_3
True	0.100	0.150	0.050	0.050	-0.100	0.100	-0.050	0.050	0.100	-0.010	10.000
Bias	0.008	0.000	0.011	0.007	0.012	0.007	0.014	0.011	0.011	0.009	1.416
RMSE	0.073	0.062	0.078	0.070	0.078	0.064	0.081	0.073	0.072	0.068	3.725
CI_L	-0.038	0.024	-0.097	-0.081	-0.249	-0.018	-0.194	-0.092	-0.039	-0.145	7.002
CI_U	0.239	0.275	0.199	0.191	0.061	0.225	0.125	0.192	0.239	0.122	21.602

of Milano-Bicocca, for providing the data used in the analysis. We recall that BTC is the first cryptocurrency that has operated digitally since 2009 with a decentralized ledger system known as blockchain. ETH, released in 2015, has a semi-decentralized network that allows creating and running smart contracts, whereby it differs from other cryptocurrencies with its unlimited supply. LTC is a clone of BTC, created in 2011. Meanwhile, XRP was created in 2012 with a different design from BTC since it has a centralized network and an un-mineable coin. Finally, BCH is an altcoin created in 2007. Especially recently, they got increasing public attention because they differentiate quite a lot from other more common assets due to their extraordinary return potential in phases of extreme price growth. We consider 1,842 daily closing prices observed over a five years period from 17 September 2017 to 02 October 2022. Log-returns of the daily closing prices are given by

$$y_{tj} = \log \frac{p_{t+1,j}}{p_{tj}}, \quad j = 1, \ldots, r, \quad t = 1, \ldots, T,$$

where p_{tj} denotes the closing price for asset j at time t. Similar data have been analyzed in Pennoni et al. (2021) through a Gaussian HMM based on discrete latent variables, to which we refer the reader for more details.

Table 2.3 presents the sample unconditional means and standard deviations of the log-returns for the five cryptocurrencies. Volatilities exhibit remarkably high values, while the average log-returns are approximately 0. Table 2.4 shows the observed linear correlations: a positive association is present for each pair, with a maximum value of 0.787 for the pair of cryptos BTC-ETH.

Table 2.3: Sample means and standard deviations (S.D.) of BTC, ETH, XRP, LTC, and BCH log-returns from 17 September 2017 to 02 October 2022.

	Cryptocurrency				
	BTC	ETH	XRP	LTC	BCH
Mean (%)	0.090	0.087	0.049	0.002	-0.073
S.D. (%)	4.166	5.271	6.451	5.659	6.585

Table 2.4: Observed correlations between log-returns of BTC, ETH, XRP, LTC, and BCH.

	BTC	ETH	XRP	LTC	BCH
BTC	1.000	-	-	-	-
ETH	0.787	1.000	-	-	-
XRP	0.560	0.653	1.000	-	-
LTC	0.765	0.823	0.644	1.000	-
BCH	0.678	0.742	0.583	0.732	1.000

2.5.2 Results

First, we assume the well-known ARMA(1,1)-GARCH(1,1) model (Engle and Bollerslev, 1986) for the marginals. Its efficacy, when combined with copula models, has been demonstrated in previous studies such as Bauwens et al. (2006) and Patton (2012). It postulates the following autoregressive equations for the conditional mean and variance of each log-return series

$$\begin{aligned} y_{tj} &= \alpha_{1j} y_{t-1,j} + \alpha_{2j} \zeta_{t-1,j} + \sqrt{\sigma^2_{tj}}\xi_{tj}, \\ \sigma^2_{tj} &= \omega_{0j} + \omega_{1j} \zeta^2_{t-1,j} + \omega_{2j} \sigma^2_{t-1,j}, \end{aligned} \tag{2.9}$$

where α_{1j}, α_{2j}, ω_{0j}, ω_{1j}, and ω_{2j}, are the ARMA(1,1)-GARCH(1,1) parameters for time-series j, $j = 1, \ldots, r$, and $\zeta_{tj} = y_{tj} - \bar{y}_{tj}$, with $\bar{y}_{tj} = \alpha_{1j} y_{t-1,j} + \alpha_{2j} \zeta_{t-1,j}$. Second, we assume that the innovations ξ_{tj} follow a skewed generalized error distribution (SGED, Theodossiou,

2015), whose density is given by

$$f(x;\phi,\kappa) = \frac{\kappa \exp\left[-\frac{1}{\kappa}\left|\frac{x+\delta_1}{\delta_2(1+\phi\,\mathrm{sign}(x+\delta_1))}\right|^{\kappa}\right]}{2\delta_2\Gamma\left(\frac{1}{\kappa}\right)}, \tag{2.10}$$

where ϕ is the skewness parameter, κ the shape parameter and

$$\begin{aligned}\delta_1 &= \frac{2^{\frac{2}{\kappa}}\delta_2\phi\Gamma\left(\frac{1}{2}+\frac{1}{\kappa}\right)}{\sqrt{\pi}},\\ \delta_2 &= \frac{\pi\left(1+3\phi^2\right)\Gamma\left(\frac{3}{\kappa}\right)-16^{\frac{1}{\kappa}}\phi^2\Gamma\left(\frac{1}{2}+\frac{1}{\kappa}\right)\Gamma\left(\frac{1}{\kappa}\right)}{\pi\Gamma\left(\frac{1}{\kappa}\right)}.\end{aligned}$$

SGED reduces to the standard Gaussian distribution when $\phi = 0$ and $\kappa = 2$ and to the Laplace distribution when $\phi = 0$ and $\kappa = 1$. It has been used previously to model univariate cryptocurrency time-series in Cerqueti et al. (2020).

Table 2.5 reports the estimated coefficients of the marginals models. Standard errors are obtained with the nonparametric block bootstrap as detailed in Section 2.3.1, considering an average block length of 127 and $B = 1,000$ bootstrap samples. Following Nasri

Table 2.5: Estimated parameters of the ARMA(1,1)-GARCH(1,1) model as in Equation (2.9). The coefficients ϕ_j and κ_j, $j = 1,\ldots,5$, refer to the skewness and shape parameters of the SGED. Standard errors (in brackets) are obtained by nonparametric block bootstrap.

	Cryptocurrency				
Parameter	BTC	ETH	XRP	LTC	BCH
α_{1j}	-0.122	-0.214	-0.021	-0.045	-0.124
	(0.009)	(0.015)	(0.003)	(0.012)	(0.014)
α_{2j}	0.038	0.113	-0.149	-0.060	0.018
	(0.003)	(0.010)	(0.008)	(0.014)	(0.009)
ω_{0j}	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ω_{1j}	0.067	0.085	0.126	0.079	0.062
	(0.006)	(0.017)	(0.021)	(0.019)	(0.013)
ω_{2j}	0.923	0.854	0.854	0.864	0.911
	(0.005)	(0.025)	(0.022)	(0.033)	(0.018)
ϕ_j	0.963	0.972	0.990	0.986	1.016
	(0.006)	(0.016)	(0.015)	(0.018)	(0.041)
κ_j	0.910	1.040	0.874	1.054	0.908
	(0.047)	(0.043)	(0.033)	(0.042)	(0.035)

and Rémillard (2019), we perform a parametric bootstrap (PB) test to evaluate the adequacy of the marginal models (Rémillard, 2011). The implementation proceeds in two steps: firstly, we generate simulated data based on the estimated marginal model; secondly, we compute the Cramér-Von Mises test statistic for the bootstraped data and compare this value with the value of the test statistic computed for the observed data in order to assess

model adequacy. Results reported in Table 2.7 show that the null hypothesis of a correct specification for the marginal distribution is never rejected at each statistical significance level. In the same table, results for the Dickey-Fuller (DF) test suggest that the null hypothesis of non-stationarity of the innovations is rejected at each significance level. We also investigate the presence of change-points in the residuals, employing the wild binary search procedure proposed by Fryzlewicz (2014). Results indicate the existence of a minimum of six change-points for all cryptocurrencies, with some displaying an even higher number.

Once we have computed the marginal pseudo-observations through the normalized ranks of the integral transformation of the innovations from the previous models, we can estimate the RSStC model, and perform model selection. The Bayesian Information Criterion (BIC, Schwarz, 1978) is commonly employed to choose a suitable number of latent states, although, it may overestimate this number. Alternatively, as demonstrated in Pohle et al. (2017), ICL provides more parsimonious results. In Table 2.7, we show the values of BIC and ICL under the RSStC model with k ranging from 1 to 4. While the BIC decreases as k increases, ICL leads us to select a model with two latent states. The maximum log-likelihood for the 2-state RSStC model is $\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) = 4,902.366$, and the total number of estimated parameters is $K = 25$.

Table 2.6: P-values of the Parametric Bootstrap (PB) and Dickey-Fuller (DF) tests.

	Cryptocurrency				
Test	BTC	ETH	XRP	LTC	BCH
PB	0.106	0.433	0.369	0.894	0.146
DF	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Table 2.7: Integrated Completed Likelihood (ICL) and Bayesian Information Criteria (BIC) computed for increasing values of the number of hidden regimes k. The minimum values are indicated in bold.

	k			
Information Criterion	1	2	3	4
ICL	-9,469.583	**-9,616.781**	-9,449.637	-9,601.758
BIC	-9,469.583	-9,887.301	-9,935.296	**-9,988.286**

Table 2.8 reports the estimated number of degrees of freedom and the determinant of the fitted matrices of dependence parameters under the RSStC model with 2 states. Notably, regimes with strong dependence exhibit a lower estimated number of degrees of freedom, indicating the prevalence of "fat-tailed" distributions in states with high correlations. This

suggests that when the correlation among crypto assets is high, joint high losses (or earnings) occur more frequently. Table 2.9 presents the matrix of the dependence parameters

Table 2.8: Estimated number of degrees of freedom ν_u, and determinant of the estimated matrices of dependence parameters under the 2-state RSStC model. Standard errors (in brackets) are obtained with nonparametric block bootstrap.

State	u =1	u =2
ν_u	6.231 (1.275)	9.416 (3.627)
$\det(\boldsymbol{R}_u)$	0.001	0.065

and Table 2.10 displays the computed Kendall's tau values using Equation (2.2). These estimates allow us to characterize each regime based on pair-specific correlations. The first regime exhibits the highest Kendall's tau values, indicating a highly correlated market state. In contrast, the second regime displays correlation values ranging from 0.324 to 0.519, suggesting a market regime with lower interdependence.

Table 2.9: Estimated dependence parameters $\rho_u^{(ij)}$ under the 2-state RSStC model. Standard errors (in brackets) are obtained with nonparametric block bootstrap.

State u =1	BTC	ETH	XRP	LTC	BCH
BTC	1.000 (0.000)	-	-	-	-
ETH	0.911 (0.016)	1.000 (0.000)	-	-	-
XRP	0.902 (0.025)	0.910 (0.020)	1.000 (0.000)	-	-
LTC	0.875 (0.034)	0.902 (0.026)	0.910 (0.034)	1.000 (0.000)	-
BCH	0.902 (0.023)	0.907 (0.022)	0.927 (0.027)	0.901 (0.035)	1.000 (0.000)
State u =2	BTC	ETH	XRP	LTC	BCH
BTC	1.000 (0.000)	-	-	-	-
ETH	0.652 (0.128)	1.000 (0.000)	-	-	-
XRP	0.667 (0.067)	0.728 (0.039)	1.000 (0.000)	-	-
LTC	0.487 (0.122)	0.613 (0.087)	0.592 (0.066)	1.000 (0.000)	-
BCH	0.569 (0.156)	0.645 (0.090)	0.672 (0.123)	0.517 (0.145)	1.000 (0.000)

Table 2.11 reports the estimated transition probability matrices under the 2-state RSStC

Table 2.10: Kendall's tau as in Equation (2.2) computed with the estimated dependence parameters $\rho_u^{(ij)}$ under the 2-state RSStC model.

State u =1	BTC	ETH	XRP	LTC	BCH
BTC	1.000	-	-	-	-
ETH	0.730	1.000	-	-	-
XRP	0.715	0.728	1.000	-	-
LTC	0.678	0.715	0.728	1.000	-
BCH	0.716	0.724	0.756	0.714	1.000
State u =2	BTC	ETH	XRP	LTC	BCH
BTC	1.000	-	-	-	-
ETH	0.452	1.000	-	-	-
XRP	0.465	0.519	1.000	-	-
LTC	0.324	0.420	0.403	1.000	-
BCH	0.385	0.446	0.469	0.346	1.000

model. We notice a general persistence in each regime: the maximum off-diagonal entry is observed from regime 2 to regime 1 (0.121). The estimated stationary distribution has probabilities $(0.579, 0.421)$.

Table 2.11: Estimated transition probabilities $\pi_{u|v}$ under the 2-state RSStC model. Standard errors (in brackets) are obtained with nonparametric block bootstrap.

State	$u=1$	$u=2$
$v=1$	0.912 (0.046)	0.088 (0.041)
$v=2$	0.121 (0.046)	0.879 (0.041)

In Table 2.12, we present the estimated state-conditional means and standard deviations for the five cryptocurrency log-returns, with the state prediction performed through the Viterbi algorithm (Viterbi, 1967). Findings reveal that the 1$^{\text{st}}$ state, which represents a regime characterized by high correlations among cryptocurrencies, is associated with negative daily log-returns. Conversely, the 2$^{\text{nd}}$ state demonstrates positive average returns. Based on these observations, we can characterize the two states as bearish and bullish market regimes. Moreover, the state-conditional standard deviations indicate high volatility in both regimes.

Figure 2.1 displays the decoded state sequence alongside the prices of BTC, ETH, XRP, LTC, and BCH. The analysis reveals distinct periods characterized by different market regimes. Initially, a bullish market regime dominates, followed by a significant presence

Table 2.12: Estimated state-conditional means and standard deviations of the five cryptocurrencies log-returns with state allocation obtained through global decoding under the 2-state RSStC model.

State 1	Mean (%)	S.D. (%)	State 2	Mean (%)	S.D. (%)
BTC	-0.363	4.121	BTC	0.744	4.147
ETH	-0.509	5.400	ETH	0.948	4.957
XRP	-0.699	5.377	XRP	1.131	7.619
LTC	-0.725	5.454	LTC	1.052	5.788
BCH	-0.902	6.013	BCH	1.125	7.169

of a bearish market regime until mid-2018. After a brief period of price increases, bearish periods becomes prominent until early 2020. Subsequently, all cryptocurrencies exhibit positive returns until late 2021, at which point a prevailing bearish trend reemerges. Notably, price increases and decreases consistently correspond to the bullish and bearish market regimes, identified solely by examining correlations. This implies that the presence of a specific market regime can be identified without relying on the first and second-order moments of cryptocurrency log-returns. Bullish and bearish regimes are visited 59.1%, 40.9% of the time, respectively, and the average sojourn times are equal to 23 days for the first state and 16 days for the second state.

Our findings corroborate the results from the previous study by Ardia et al. (2019), in which the authors use a 2-state Markov-switching GARCH model fitted on univariate BTC time-series. Their model employs a fat-tailed distribution across two regimes identified by low and high unconditional volatilities which exhibit strong persistence. Koki et al. (2022) fit a range of HMMs to the log-returns of three cryptocurrencies, namely BTC, ETH, and XRP, with different numbers of states. They determine that a 4-state model provided the most accurate forecasting results among all considered model specifications. Nonetheless, the statistical properties of the hidden states exhibit differences among the three cryptocurrencies, making the interpretation of these latent states as distinct economic regimes a challenging task.

2.5.3 Comparative analysis

We compare our proposal with the estimates obtained with the basic HMM (Zucchini et al., 2017). We also show the forecast performance of both models and a benchmark model, the MRW process.

The basic HMM assumes a conditional Gaussian distribution and it is generally not robust for analyzing extreme events usually observed in financial data. Cryptocurrency markets are characterized by heavy-tailed distributions that lead to frequent extreme price

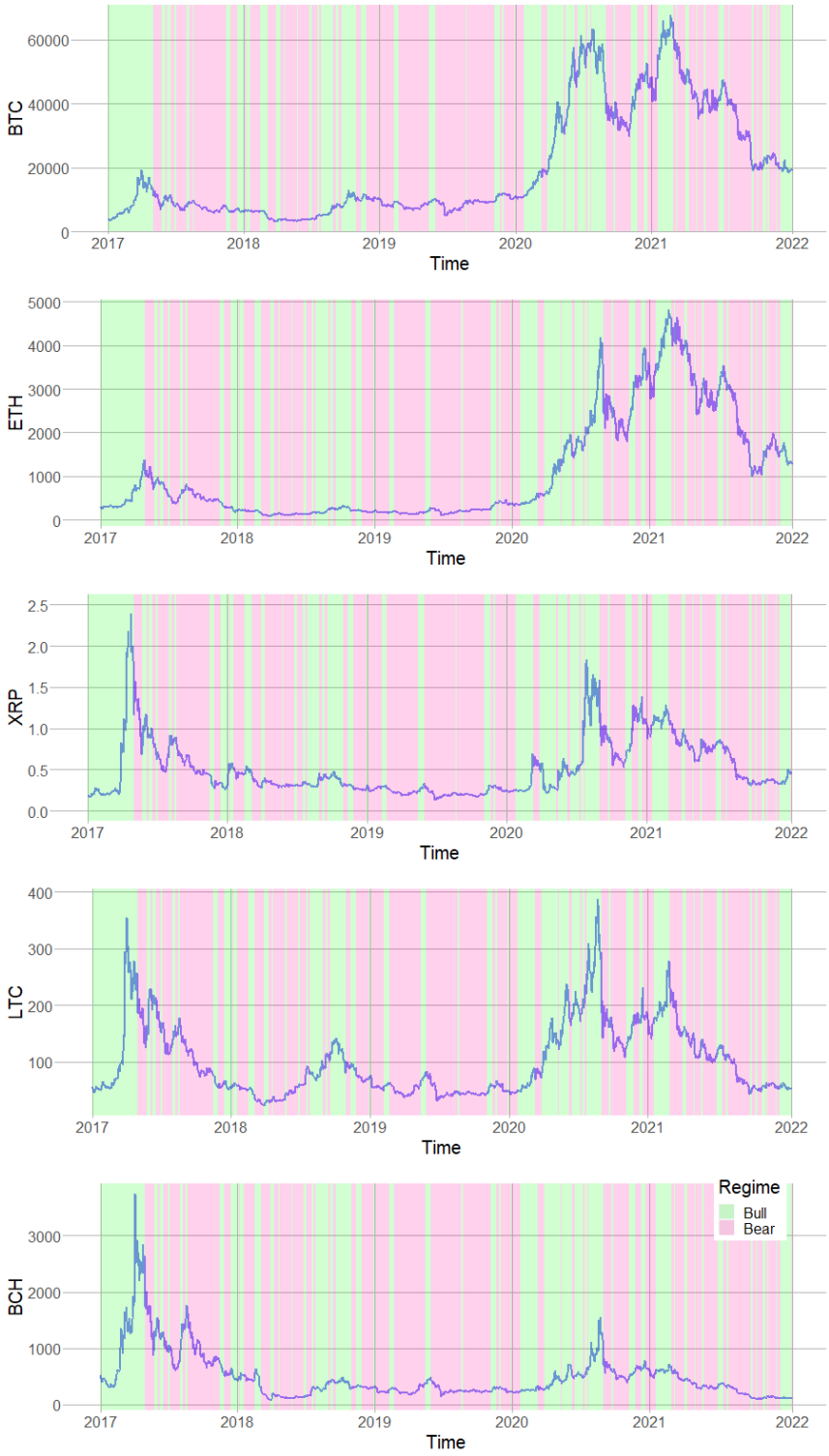


Figure 2.1: Observed prices of BTC, ETH, XRP, LTC, and BCH (17 September 2017 - 02 October 2022) with the global decoding state sequence highlighted in red for state 1 (bearish market) and green for state 2 (bullish market).

movements, whether positive or negative, more than in traditional financial markets. In this regard, the Student-t copula accommodates heavy-tailed distributions, allowing us to model the idiosyncrasies of cryptocurrency log-returns. Thus, the proposed RSStC model may appropriately represents the observed underlying trends of the cryptocurrency log-returns.

Estimation of the parameters of the HMM has been performed with the routines provided within the `R` package **RcppHMM** (Ardenas-Ovando et al., 2017); for an alternative HMM formulation, the **LMest** package (Bartolucci et al., 2017) can be used. Model selection performed with both BIC and ICL criteria suggests a HMM with 6 regimes. In the following, we show some results obtained with both 2- and 6-state HMMs, denoted as HMM-2, and HMM-6, respectively.

We note that the self-transition probabilities estimated under the HMM-2 are 0.633 and 0.873, respectively, and those of the HMM-6 range in the interval (0.425, 0.673). Notably, as illustrated in the previous section and shown in Table 2.11, the RSStC model exhibits higher values. Such a model is sometimes preferable to make profitable investment decisions since asset allocation is more stable and the portfolio turnover is low. In fact, disposing of a model with more stable regimes allows us to avoid frequent reallocation or trading in response to short-term market fluctuations. This strategy implies reduced transaction costs and tax-efficient investments strategies, as also noticed in Nystrup et al. (2020).

Forecasts are implemented through a rolling window approach for each model: it consists in dividing the available data into overlapping windows, each consisting of 1,500 observations. We estimate parameters using each window and employ the estimates to generate one-step-ahead forecasts for the log-returns of the five cryptocurrencies. In more details, we obtain forecasts for the RSStC model according to the following steps, as suggested in Simard and Rémillard (2015):

1. Estimate the RSStC model as explained in Sections 2.2 and 2.3.
2. Simulate pseudo-marginals from the fitted copula model. This involves generating random samples from the copula function corresponding to the estimated dependence structure. In particular, we consider $1,000$ observations.
3. Transform the generated uniform samples into the desired marginal distributions using the inverse of the estimated marginal cumulative distribution functions.
4. Use data obtained at the previous step to forecast future values of each variable in the time-series using sample mean, and estimate prediction intervals using sample quantiles.

To estimate HMM and MRW forecasts, we adopt a similar approach. We utilize the estimated parameters for the HMM to simulate portfolio realizations for the one-step-ahead

observation. This process involves generating latent states based on the HMM's predictive distribution. These latent states are then used to create corresponding portfolio returns for the next period. Similarly, for the MRW process, we use the estimated parameters to simulate future portfolio returns.

We evaluate forecast quality through two different metrics. We employ the RMSE and the percentage of correct sign predictions (CSP). RMSE quantifies the accuracy between true and forecasted values by calculating the average of squared differences. CSP, on the other hand, measures the frequency with which we accurately forecast the sign of returns. We present the results of our evaluation in Table 2.13.

Table 2.13: RMSE between true and forecasted values of the five cryptocurrencies and percentage CSP obtained under the 2-state RSStC, HMM-2, HMM-6 and MRW models.

	RSStC	HMM-2	HMM-6	MRW
RMSE	**0.061**	0.065	0.067	0.093
CSP (%)	**53.26**	49.03	50.73	50.67

As Timmermann (2018) wisely noted, "*detecting breaks in financial forecasting models is a formidable task, and transforming such evidence into more accurate forecasts is even more challenging*". Our approach acknowledges this inherent complexity, positioning it as a practical solution for decision-making processes, even without aiming for extraordinary predictive accuracy. In fact, in terms of forecasting performance, the RSStC model achieves lower RMSE compared to traditional models, thus providing enhanced accuracy in predicting cryptocurrency returns which can results particularly useful in case of future financial crisis events.

We also compare the models using the Diebold-Mariano test (Diebold and Mariano, 2002) and Model Confidence Set (MCS) procedure of Hansen et al. (2011). The first test specifically compares the forecast accuracy of two models, typically testing whether the difference in forecast performance is statistically significant. The MCS, instead, constructs a confidence set that includes models that are statistically indistinguishable from the best-performing model. The Diebold-Mariano test reveals no statistically significant difference among the squared error estimates of the three models. However, according to the MCS, the RSStC emerges as the top predictive model for LTC, XRP, and BCH log-returns. In contrast, for BTC and ETH log-returns, the MRW ranks first.

We additionally evaluate the ability of the three models to cover severe losses computing a popular risk measure like the Value-at-Risk (VaR, Duffie and Pan, 1997). We evaluate whether the VaR forecasts adequately capture the actual proportion of losses surpassing the estimated VaR level throught the conditional coverage test of Christoffersen (1998). The null hypothesis typically posits that the VaR model correctly captures the proportion

or frequency of actual losses exceeding the VaR threshold at the specified confidence level. The findings indicate that, at the 1% significance level, the null hypothesis is not rejected for RSStC and HMM VaR forecasts, while being rejected for the MRW model.

2.6 Discussion

In this Chapter, we provide three main contributions to the existing literature on regime switching copula models. First, we generalize the regime switching Student-t copula model to the multivariate case. Second, we propose a novel maximum likelihood estimation procedure for multivariate regime switching Student-t copula models through a two-step method, initially estimating the dependence matrix and then the number of degrees of freedom. Third, we analyze the joint distribution of the log-returns of five cryptocurrencies during the period 2017-2022 with the proposed model.

Our simulation studies demonstrate the good finite sample properties of the proposed estimators, highlighting its ability to accurately detect the true number of degrees of freedom. This capability is particularly pronounced in situations where this number is low, denoting a process with fat tails.

By analyzing five years of time-series data encompassing the log-returns of Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin Cash, we show the suitability of a 2-state regime switching Student-t copula model for detecting market trends. Through the application of the Viterbi algorithm to decode state sequences, we can effectively distinguish between bullish and bearish market phases. Notably, bearish periods in financial markets correspond to increasing correlations among assets. In comparison to two commonly used basic models, our proposed approach is more robust and yields slightly improved forecasting results. These advantages translate into the potential for making more profitable investment decisions and implementing portfolio trading strategies. The estimated allocation into each regime, as determined through global decoding, plays a crucial role in achieving these benefits. Given the ongoing growth of these cryptocurrencies and the presence of co-integration and dynamic interdependencies between them, the ability to detect signals of market dynamics is of primary importance, not only for optimizing investment strategies but also for identifying early warnings of potential financial crises.

As lines for future research, we highlight that it would be of interest to model marginals distributions with a Markov process along with the joint distribution. Moreover, an interesting area of investigation lies in the use of skewed Student-t copulas (Bauwens and Laurent, 2005). These copulas present a promising approach to capture asymmetries in dynamic behaviors.

Appendices

2.A EM algorithm implementation

Let us denote with $\boldsymbol{\theta}$ the full vector of parameters and with $\ell^*\left(\boldsymbol{\theta} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T)\right)$ the complete-data log-likelihood as in Equation (2.3). Starting from an initial parameter vector $\boldsymbol{\theta}^{(0)}$, at the m-iteration the algorithm performs the following steps:

- **E-step**: compute the posterior expected value of the indicator variables w_{tu}, z_{tuv}, given by the quantities

$$\begin{aligned} \hat{w}_{tu} &= P(u_t = u \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T), \\ \hat{z}_{tuv} &= P(u_{t-1} = u, u_t = v \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T,), \end{aligned}$$

for $t = 1, \ldots, T$, and for all $u, v = 1, \ldots, k$. Let us define (see also Remillard (2013); Nasri et al. (2020))

$$\bar{\eta}_t(u) = P(u_t = u \mid \boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_T), \quad t = 1, \ldots, T,$$

$$\eta_t(u) = P(u_t = u \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t), \quad t = 2, \ldots, T,$$

where the conditioning argument disappears from the first expression for $t = T$. The above quantities are initialized with

$$\bar{\eta}_T(u) = 1/k, \quad \eta_1(u) = \frac{\lambda_u c(\boldsymbol{y}_1; \boldsymbol{R}_u, \nu_u)}{\sum_{v=1}^k \pi_v c(\boldsymbol{y}_1; \boldsymbol{R}_v, \nu_v)}, \quad u = 1, \ldots, k,$$

and are computed recursively (Baum and Petrie, 1966; Welch, 2003; Nasri et al., 2020) through

$$\eta_t(u) = \frac{c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u) \sum_{v=1}^k \eta_{t-1}(v) \pi_{u|v}}{\sum_{a=1}^k c(\boldsymbol{y}_t; \boldsymbol{R}_a, \nu_a) \sum_{v=1}^k \eta_{t-1}(v) \pi_{a|v}}, \quad t = 2, \ldots, T,$$

$$\bar{\eta}_t(u) = \frac{\sum_{v=1}^k \bar{\eta}_{t+1}(v) \pi_{v|u} c(\boldsymbol{y}_{t+1}; \boldsymbol{R}_v, \nu_v)}{\sum_{a=1}^k \sum_{v=1}^k \bar{\eta}_{t+1}(v) \pi_{v|a} c(\boldsymbol{y}_{t+1}; \boldsymbol{R}_v, \nu_v)}, \quad t = 1, \ldots, T-1,$$

to be evaluated in reverse order. From the previous expressions we obtain $\hat{w}_{tu}$ and $\hat{z}_{tvu}$ as

$$\hat{w}_{tu} = \frac{\eta_t(u) \bar{\eta}_t(u)}{\sum_{a=1}^k \eta_t(a) \bar{\eta}_t(a)}, \quad t = 1, \ldots, T, \; u = 1, \ldots, k,$$

$$\hat{z}_{tuv} = \frac{\pi_{v|u} \eta_{t-1}(u) \bar{\eta}_t(v) c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u)}{\sum_{a=1}^k \sum_{b=1}^k \pi_{b|a} \eta_{t-1}(a) \bar{\eta}_t(b) c(\boldsymbol{y}_t; \boldsymbol{R}_b, \nu_b)}, \quad t = 2, \ldots, T, \; u, v = 1, \ldots, k.$$

The expected value of the complete-data log-likelihood is then obtained by substituting $\hat{w}_{tu}$ and $\hat{z}_{tuv}$ to w_{tu} and z_{tuv}, in Equation (2.3). This expected value is denoted by $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m-1)})$ where $\boldsymbol{\theta}^{(m-1)}$ is the vector of parameters provided by the previous M-step, on the basis of which $\hat{w}_{tu}$ and $\hat{z}_{tuv}$ are computed.

- **M-step**. The new parameter vector $\boldsymbol{\theta}^{(m)}$ is obtained as $\text{argmax}_{\boldsymbol{\theta}}\, \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m-1)})$. Parameters λ_u and $\pi_{v|u}$ are updated by using formulas in (2.4) and (2.5). Following Trede (2020), the updated values of the remaining parameters, $\boldsymbol{R}_u^{(m)}$ and $\nu_u^{(m)}$, are obtained as follows: for a given u, $u = 1, \ldots, k$, we maximize

$$\sum_{t=1}^{T} \hat{w}_{tu} \log c(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u),$$

subject to the restriction that $\boldsymbol{R}_u$ is symmetric, positive definite, and with all diagonal elements equal to 1. The Lagrangians are the following

$$\mathcal{L}(\boldsymbol{R}_u \mid \nu_u) = \sum_{t=1}^{T} \hat{w}_{tu} \log c\left(\boldsymbol{y}_t; \boldsymbol{R}_u, \nu_u\right) + \sum_{j=1}^{r} \mu_j \left(\rho_u^{(jj)} - 1\right),$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)'$ being the Lagrange multipliers. Setting the first derivative with respect to $\boldsymbol{R}_u$ equal to 0 turns to

$$\frac{\partial \mathcal{L}(\boldsymbol{R}_u \mid \nu_u)}{\partial \boldsymbol{R}_u} = -\frac{\sum_t \hat{w}_{tu}}{2} \boldsymbol{R}_u^{-1} + \frac{\nu_u + r}{2\nu_u} \sum_{t=1}^{T} \hat{w}_{tu} \boldsymbol{R}_u^{-1} \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{R}_u^{-1} \left(1 + \frac{1}{v} \boldsymbol{x}_t' \boldsymbol{R}_u^{-1} \boldsymbol{x}_t\right)^{-1} + \boldsymbol{M} = \boldsymbol{0},$$

where we denoted with $\boldsymbol{M} = \text{diag}(\boldsymbol{\mu})$ the matrix with diagonal elements equal to $\mu_1, \ldots, \mu_r$, and zero in all other positions and with $\boldsymbol{x}_t$ the vector with components $T_{1,\nu_u}^{-1}(y_{tj})$, with $j = 1, \ldots, r$. Multiplying both sides by $\boldsymbol{R}_u$ gives

$$\boldsymbol{R}_u = \frac{\nu_u + r}{\nu_u \sum_{t=1}^{T} \hat{w}_{tu}} \sum_{t=1}^{T} \hat{w}_{tu} \boldsymbol{x}_t \boldsymbol{x}_t' \left(1 + \frac{1}{\nu_u} \boldsymbol{x}_t' \boldsymbol{R}_u^{-1} \boldsymbol{x}_t\right)^{-1} + \frac{2}{\sum_t \hat{w}_{tu}} \boldsymbol{R}_u \boldsymbol{M} \boldsymbol{R}_u.$$

Let denote the first term of the previous expression with

$$\boldsymbol{A} = \frac{\nu_u + r}{\nu_u \sum_t \hat{w}_{tu}} \sum_{t=1}^{T} \hat{w}_{tu} \boldsymbol{x}_t \boldsymbol{x}_t' \left(1 + \frac{1}{\nu_u} \boldsymbol{x}_t' \boldsymbol{R}_u^{-1} \boldsymbol{x}_t\right)^{-1},$$

and let $\boldsymbol{a}$ be the vector of diagonal elements of $\boldsymbol{A}$. The Lagrange multipliers $\boldsymbol{\mu}$ satisfy the equations

$$\frac{2}{\sum_{t=1}^{T} \hat{w}_{tu}} (\boldsymbol{R}_u \circ \boldsymbol{R}_u) \boldsymbol{\mu} = \boldsymbol{1} - \boldsymbol{a},$$

with $\boldsymbol{R}_u \circ \boldsymbol{R}_u$ being the matrix of squared elements of $\boldsymbol{R}_u$. In order to satisfy the restrictions, the Lagrange multipliers are

$$\boldsymbol{\mu} = \sum_{t=1}^{T} \hat{w}_{tu} (\boldsymbol{R}_u \circ \boldsymbol{R}_u)^{-1} (\mathbf{1} - \boldsymbol{a})/2,$$

which, substituted in the equation for $\boldsymbol{R}_u$, yields to

$$\boldsymbol{R}_u = \boldsymbol{A} + \boldsymbol{R}_u \text{diag}\left[(\boldsymbol{R}_u \circ \boldsymbol{R}_u)^{-1} (\mathbf{1} - \boldsymbol{a}) \right] \boldsymbol{R}_u.$$

The above Equation cannot be solved analytically and we consider the iterative solution as in Equation (2.7). The estimate for ν_u is obtained by solving the optimization problem in Equation (2.8).

In the application presented in Section 2.5, the number of latent states k is selected according to the ICL criterion (Biernacki et al., 2000), defined as

$$\text{ICL} = -2 \log \ell^*(\hat{\boldsymbol{\theta}} \mid (\boldsymbol{y}_1, u_1), \ldots, (\boldsymbol{y}_T, u_T)) + K \log(T), \tag{2.11}$$

being $\hat{\boldsymbol{\theta}}$ the vector of estimated RSStC parameters and K the number of free parameters, computed as $K = (k-1) + k(k-1) + k(r(r+1)/2) + k$. As suggested in Pohle et al. (2017), the unknown sequence $u_1, \ldots, u_T$ may be replaced with the decoded time-series $\hat{u}_1, \ldots, \hat{u}_T$ obtained applying the Viterbi algorithm to the posterior probabilities estimated with the selected RSStC model.

2.B Complete simulation results

In this Appendix, we report the simulation results regarding the 2-state RSStC model, and the complete results for the 2- and for the 3-state model when varying the number of observations T or the number of assets r.

Two state regime switching Student-t copula model

We generate data from a 2-state RSStC model and we simulate $B = 1,000$ samples of dimension $r = 5$, each with a total number of observations T =1,500. We set the vector of initial probabilities $\boldsymbol{\lambda} = (1/2, 1/2)'$ so that each state is equally likely, we fix the transition matrix assumed as follows

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.800 & 0.200 \\ 0.200 & 0.800 \end{bmatrix},$$

the matrices of dependence parameters as

$$\boldsymbol{R}_1 = \begin{bmatrix} 1.000 & - & - & - & - \\ 0.900 & 1.000 & - & - & - \\ 0.900 & 0.900 & 1.000 & - & - \\ 0.900 & 0.900 & 0.900 & 1.000 & - \\ 0.900 & 0.900 & 0.900 & 0.900 & 1.000 \end{bmatrix}, \quad \boldsymbol{R}_2 = \begin{bmatrix} 1.000 & - & - & - & - \\ 0.200 & 1.000 & - & - & - \\ 0.000 & 0.000 & 1.000 & - & - \\ 0.100 & 0.200 & 0.100 & 1.000 & - \\ 0.000 & 0.200 & 0.200 & 0.000 & 1.000 \end{bmatrix},$$

and the state-specific numbers of degrees of freedom are fixed as $\nu_1 = 5$ and $\nu_2 = 15$, respectively. We simulate a turbulent market scenario with high asset correlation and fat tails, and a more stable market scenario with low dependencies and a higher degree of freedom. We evaluate the results using the criteria outlined in Section 2.4.

Table 2.14 presents the true value, bias, RMSE, and CI for each parameter. The maximum absolute bias for the initial probabilities is 0.033, and for the transition probabilities is approaching zero. The maximum RMSE is 0.499 for the initial probabilities and 0.018 for the transition probabilities. CIs for the transition probabilities are narrow and centered around the true values, those of the initial probabilities reflect the unitary nature of the maximum likelihood estimator for these parameters (Zucchini et al., 2017).

Table 2.14: Simulation results for the 2-state RSStC model: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{2\|1}$	$\pi_{2\|2}$
True	0.500	0.500	0.800	0.200	0.200	0.800
Bias	-0.033	0.033	0.000	0.000	0.000	0.000
RMSE	0.499	0.499	0.018	0.018	0.017	0.017
CI_L	0.000	0.000	0.764	0.168	0.168	0.764
CI_U	1.000	1.000	0.832	0.236	0.236	0.832

Table 2.15 presents simulation results for the dependence parameters $\rho_u^{(ij)}$ and the number of degrees of freedom ν_u. Regarding the dependence parameters, the maximum absolute bias is 0.002, and the highest RMSE occurs for $\rho_2^{(15)}$ and $\rho_2^{(45)}$, with a value of 0.041. The CIs are narrower in the first state, indicating a more accurate estimation of high correlations. In terms of the number of degrees of freedom, we observe that as ν_u decreases, the absolute bias and RMSE decrease while the CI narrows.

Increasing the series length and the number of assets

In this simulated scenario, we investigate the RMSE by varying the series length (T) and the number of assets (r). We simulate data from 2- and 3-state RSStC models using the values of the parameters employed for the simulated scenarios presented in Sections 2.B and 2.4.1.

Table 2.15: Simulation results for the 2-state RSStC model: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	5.000
Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	-0.001	0.000	-0.157
RMSE	0.008	0.009	0.008	0.008	0.009	0.008	0.009	0.009	0.009	0.008	0.916
CI_L	0.882	0.881	0.883	0.882	0.880	0.882	0.881	0.882	0.882	0.882	3.724
CI_U	0.915	0.915	0.915	0.915	0.915	0.915	0.915	0.916	0.916	0.915	7.159
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.200	0.000	0.100	0.000	0.000	0.200	0.200	0.100	0.200	0.000	15.000
Bias	0.002	0.000	0.001	0.001	0.002	0.001	0.001	0.001	-0.001	0.002	0.668
RMSE	0.038	0.040	0.040	0.041	0.040	0.037	0.037	0.039	0.038	0.041	3.476
CI_L	0.128	-0.077	0.024	-0.080	-0.077	0.132	0.125	0.022	0.123	-0.080	10.480
CI_U	0.274	0.083	0.175	0.077	0.081	0.270	0.274	0.180	0.275	0.075	23.502

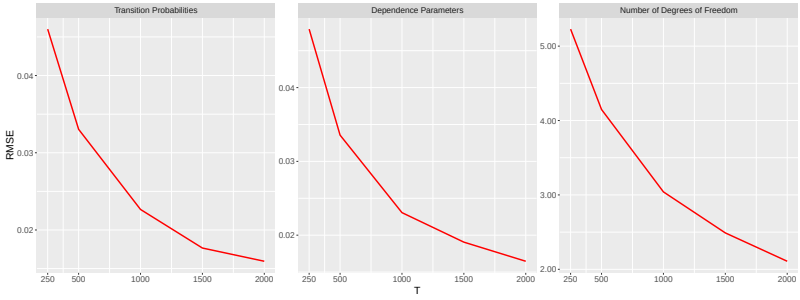
For each simulated sample, we calculate the average RMSE over $B = 1,000$ samples for each group of parameters, including initial probabilities, transition probabilities, number of degrees of freedom, and dependence parameters. We consider five different series of lengths $T = 250$, 500, 1,000, 1,500, 2,000, referred to $r = 5$ assets.

We summarize the results in Figure 2.2, illustrating that the average RMSE decreases rapidly as the series length increases for all parameters. We omit the average RMSE for initial probabilities as it remains around 0.5 up to the fourth decimal digit.

Then, we increase the number of assets while keeping the series length fixed at $T = 1,000$. We consider three plausible values for r, namely 2, 5, and 10, and we simulate $B = 1,000$ samples from the 2- and 3-state RSStC models. In the 2-state model, the initial and transition probabilities, as well as the vector of degrees of freedom, remain unchanged from the first simulation study. However, the dependence parameters are kept identical for all pairs of observations. Specifically, we use $\rho_1^{(ij)} = 0.9$ for the first regime and $\rho_2^{(ij)} = 0.1$ for the second regime, for $i \neq j$. For the 3-state RSStC model, we follow a similar procedure. The initial probabilities, transition probabilities, and number of degrees of freedom remain the same as in the second simulation study. The dependence parameters are set as $\rho_1^{(ij)} = 0.9$, $\rho_2^{(ij)} = 0.5$, and $\rho_3^{(ij)} = 0.1$, for $i \neq j$. In Figure 2.3, we observe that the average RMSE for the dependence parameters increases when the number of assets is higher than 2. However, the RMSE for the number of degrees of freedom and the transition probabilities decreases as the number of marginals increases.

In the following, we present comprehensive results of the simulation studies conducted

(a) 2-state RSStC model



(b) 3-state RSStC model

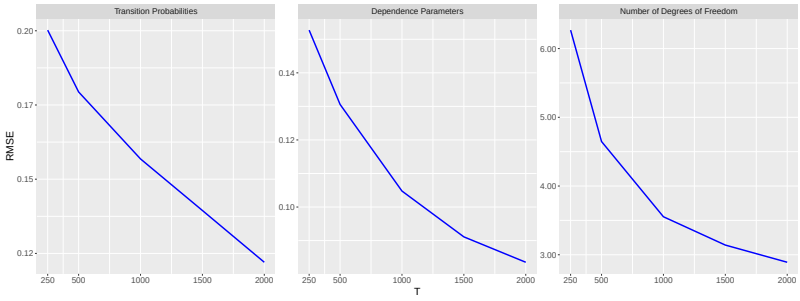
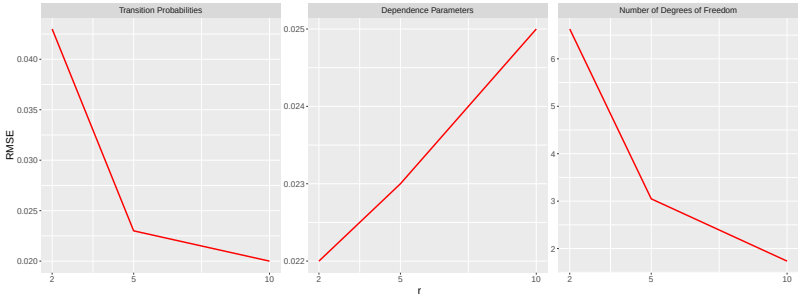


Figure 2.2: Average RMSE for transition probabilities, dependence parameters, and number of degrees of freedom in the 2-state (a) and 3-state (b) RSStC models, with series length (T) varying from 250 to 2,000 and $r = 5$.

(a) 2-state RSStC model



(b) 3-state RSStC model

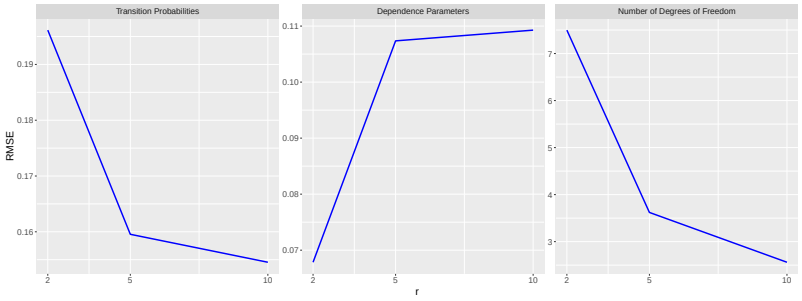


Figure 2.3: Average RMSE for transition probabilities, dependence parameters, and number of degrees of freedom in the 2-state (a) and 3-state (b) RSStC models, with number of assets (r) varying from 2 to 10 and $T = 1,500$.

in Section 4. Specifically, Tables 2.16 to 2.21 report the results for the 2-state model for $T = 250, 500$, and $T = 1,000$. Additionally, Tables 2.22 to 2.27 present the results for the 3-state model, for $T = 250, 500$, and $T = 1,000$.

Table 2.16: Simulation results for the 2-state RSStC model with $T = 250$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{2\|1}$	$\pi_{2\|2}$
True	0.500	0.500	0.800	0.200	0.200	0.800
Bias	-0.034	-0.034	0.002	-0.002	-0.003	0.003
RMSE	0.500	0.500	0.046	0.046	0.046	0.046
CI_L	0.000	0.000	0.690	0.120	0.124	0.698
CI_U	1.000	1.000	0.880	0.310	0.302	0.876

Table 2.17: Simulation results for the 2-state RSStC model with $T = 250$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	5.000
Bias	0.002	0.002	0.002	0.002	0.002	0.001	0.003	0.002	0.002	0.001	-1.142
RMSE	0.022	0.023	0.022	0.023	0.022	0.022	0.022	0.023	0.023	0.022	3.520
CI_L	0.848	0.849	0.845	0.849	0.849	0.846	0.852	0.846	0.849	0.848	2.772
CI_U	0.935	0.934	0.932	0.935	0.933	0.934	0.932	0.936	0.934	0.933	14.605
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.200	0.000	0.100	0.000	0.000	0.200	0.200	0.100	0.200	0.000	15.000
Bias	-0.002	-0.003	-0.006	-0.001	-0.001	0.000	-0.001	-0.003	-0.001	-0.000	-3.090
RMSE	0.093	0.104	0.098	0.101	0.098	0.093	0.095	0.099	0.091	0.102	6.937
CI_L	0.004	-0.209	-0.093	-0.195	-0.192	0.016	0.010	-0.103	0.018	-0.197	7.468
CI_U	0.364	0.198	0.286	0.204	0.187	0.378	0.387	0.281	0.378	0.198	25.000

Table 2.18: Simulation results for the 2-state RSStC model with $T = 500$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{2\|1}$	$\pi_{2\|2}$
True	0.500	0.500	0.800	0.200	0.200	0.800
Bias	-0.023	0.023	0.002	-0.002	-0.002	0.002
RMSE	0.500	0.500	0.034	0.034	0.032	0.032
CI_L	0.000	0.000	0.728	0.142	0.141	0.733
CI_U	1.000	1.000	0.858	0.272	0.267	0.859

Table 2.19: Simulation results for the 2-state RSStC model with $T = 500$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	5.000
Bias	0.001	0.001	0.000	0.001	0.001	-0.000	0.001	0.000	0.000	0.000	-0.631
RMSE	0.015	0.016	0.015	0.016	0.016	0.015	0.015	0.016	0.016	0.016	2.183
CI_L	0.865	0.865	0.867	0.863	0.867	0.866	0.865	0.866	0.864	0.863	3.094
CI_U	0.925	0.927	0.927	0.924	0.926	0.927	0.925	0.927	0.927	0.926	11.292
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.200	0.000	0.100	0.000	0.000	0.200	0.200	0.100	0.200	0.000	15.000
Bias	0.000	0.001	-0.001	0.001	0.002	-0.003	-0.001	0.000	-0.003	-0.001	-2.525
RMSE	0.067	0.069	0.070	0.071	0.071	0.065	0.065	0.069	0.065	0.070	6.113
CI_L	0.067	-0.138	-0.035	-0.129	-0.139	0.073	0.073	-0.041	0.071	-0.131	8.615
CI_U	0.330	0.128	0.235	0.142	0.135	0.326	0.329	0.233	0.318	0.132	25.000

Table 2.20: Simulation results for the 2-state RSStC model with $T = 1,000$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{2\|1}$	$\pi_{2\|2}$
True	0.500	0.500	0.800	0.200	0.200	0.800
Bias	-0.020	0.020	0.000	0.000	-0.001	0.001
RMSE	0.500	0.500	0.023	0.023	0.022	0.022
CI_L	0.000	0.000	0.751	0.158	0.161	0.753
CI_U	1.000	1.000	0.842	0.249	0.247	0.839

Table 2.21: Simulation results for the 2-state RSStC model with $T = 1,000$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	5.000
Bias	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.000	0.000	0.000	-0.242
RMSE	0.010	0.011	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.01	1.147
CI_L	0.878	0.876	0.880	0.878	0.879	0.878	0.879	0.877	0.878	0.878	3.581
CI_U	0.919	0.919	0.919	0.918	0.919	0.918	0.919	0.918	0.919	0.918	7.883
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.200	0.000	0.100	0.000	0.000	0.200	0.200	0.100	0.200	0.000	15.000
Bias	0.000	0.002	0.000	-0.001	0.001	0.000	-0.004	-0.001	-0.001	-0.001	-1.588
RMSE	0.047	0.048	0.048	0.049	0.049	0.046	0.045	0.047	0.044	0.05	4.935
CI_L	0.109	-0.093	0.004	-0.085	-0.097	0.109	0.114	0.008	0.111	-0.098	9.719
CI_U	0.286	0.093	0.194	0.094	0.093	0.295	0.294	0.195	0.284	0.103	25.000

Table 2.22: Simulation results for the 3-state RSStC model with $T = 250$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	λ_3	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|3}$	$\pi_{2\|1}$	$\pi_{2\|2}$	$\pi_{2\|3}$	$\pi_{3\|1}$	$\pi_{3\|2}$	$\pi_{3\|3}$
True	0.333	0.333	0.333	0.700	0.200	0.100	0.200	0.700	0.100	0.100	0.200	0.700
Bias	-0.038	0.055	-0.017	0.041	0.020	-0.060	-0.054	0.348	-0.293	-0.033	-0.068	0.102
RMSE	0.483	0.445	0.473	0.144	0.129	0.121	0.169	0.406	0.353	0.088	0.183	0.210
CI_L	0.000	0.000	0.000	0.255	0.000	0.000	0.000	0.000	0.064	0.000	0.024	0.175
CI_U	1.000	1.000	1.000	0.831	0.477	0.385	0.614	0.755	0.821	0.326	0.704	0.856

Table 2.23: Simulation results for the 3-state RSStC model with $T = 250$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.700	0.800	0.800	0.750	0.900	0.800	0.700	0.800	0.800	3.000
Bias	0.030	0.021	0.022	0.027	0.028	0.024	0.028	0.021	0.023	0.025	-0.303
RMSE	0.099	0.062	0.074	0.085	0.088	0.075	0.085	0.063	0.065	0.077	1.974
CI_L	0.740	0.767	0.771	0.754	0.756	0.748	0.733	0.770	0.769	0.744	2.000
CI_U	0.955	0.954	0.957	0.953	0.952	0.954	0.956	0.952	0.953	0.953	8.99
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.500	0.300	0.500	0.400	0.400	0.400	0.500	0.400	0.500	0.300	7.000
Bias	-0.048	0.006	-0.003	-0.020	-0.024	-0.023	-0.041	-0.003	-0.019	-0.020	-4.188
RMSE	0.279	0.285	0.282	0.301	0.301	0.249	0.285	0.274	0.259	0.299	8.231
CI_L	-0.170	-0.159	-0.244	-0.267	-0.275	-0.049	-0.168	-0.165	-0.096	-0.241	2.684
CI_U	0.899	0.904	0.898	0.898	0.906	0.908	0.897	0.903	0.901	0.898	25.000
State 3	$\rho_3^{(12)}$	$\rho_3^{(13)}$	$\rho_3^{(14)}$	$\rho_3^{(15)}$	$\rho_3^{(23)}$	$\rho_3^{(24)}$	$\rho_3^{(25)}$	$\rho_3^{(34)}$	$\rho_3^{(35)}$	$\rho_3^{(45)}$	ν_1
True	0.100	0.150	0.050	0.050	-0.100	0.100	-0.050	0.050	0.100	-0.010	15.000
Bias	-0.039	-0.181	-0.128	-0.135	-0.129	-0.092	-0.056	-0.141	-0.088	-0.135	-6.185
RMSE	0.172	0.269	0.226	0.229	0.234	0.191	0.178	0.226	0.192	0.225	8.600
CI_L	-0.137	-0.250	-0.202	-0.255	-0.281	-0.072	-0.116	-0.146	-0.089	-0.254	7.338
CI_U	0.531	0.548	0.544	0.472	0.483	0.577	0.559	0.546	0.579	0.452	25.000

Table 2.24: Simulation results for the 3-state RSStC model with $T = 500$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	λ_3	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|3}$	$\pi_{2\|1}$	$\pi_{2\|2}$	$\pi_{2\|3}$	$\pi_{3\|1}$	$\pi_{3\|2}$	$\pi_{3\|3}$
True	0.333	0.333	0.333	0.700	0.200	0.100	0.200	0.700	0.100	0.100	0.200	0.700
Bias	-0.026	0.074	-0.048	0.037	0.029	-0.066	-0.061	0.330	-0.269	-0.022	-0.040	0.062
RMSE	0.478	0.438	0.485	0.130	0.106	0.118	0.158	0.387	0.325	0.063	0.155	0.172
CI_L	0.000	0.000	0.000	0.278	0.013	0.017	0.033	0.000	0.069	0.012	0.035	0.251
CI_U	1.000	1.000	1.000	0.797	0.430	0.357	0.573	0.772	0.790	0.245	0.609	0.850

Table 2.25: Simulation results for the 3-state RSStC model with $T = 500$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.700	0.800	0.800	0.750	0.900	0.800	0.700	0.800	0.800	3.000
Bias	0.019	0.016	0.016	0.017	0.019	0.014	0.017	0.017	0.016	0.017	-0.015
RMSE	0.079	0.063	0.071	0.070	0.099	0.057	0.059	0.074	0.063	0.075	1.120
CI_L	0.808	0.819	0.813	0.812	0.818	0.822	0.813	0.823	0.808	0.807	2.000
CI_U	0.942	0.941	0.944	0.943	0.944	0.943	0.944	0.936	0.943	0.942	5.998
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.500	0.300	0.500	0.400	0.400	0.400	0.500	0.400	0.500	0.300	7.000
Bias	-0.096	-0.056	-0.055	-0.072	-0.078	-0.063	-0.087	-0.058	-0.054	-0.072	-1.865
RMSE	0.246	0.239	0.231	0.264	0.257	0.216	0.248	0.229	0.224	0.259	5.295
CI_L	0.043	-0.039	0.027	-0.070	-0.079	0.086	0.024	0.061	0.108	-0.055	2.677
CI_U	0.900	0.908	0.903	0.902	0.890	0.896	0.905	0.910	0.910	0.900	25.000
State 3	$\rho_3^{(12)}$	$\rho_3^{(13)}$	$\rho_3^{(14)}$	$\rho_3^{(15)}$	$\rho_3^{(23)}$	$\rho_3^{(24)}$	$\rho_3^{(25)}$	$\rho_3^{(34)}$	$\rho_3^{(35)}$	$\rho_3^{(45)}$	ν_1
True	0.100	0.150	0.050	0.050	-0.100	0.100	-0.050	0.050	0.100	-0.010	15.000
Bias	-0.042	-0.171	-0.131	-0.128	-0.120	-0.094	-0.045	-0.128	-0.092	-0.128	-3.577
RMSE	0.133	0.231	0.192	0.197	0.191	0.156	0.135	0.195	0.152	0.192	7.527
CI_L	-0.029	-0.161	-0.088	-0.189	-0.211	0.022	-0.054	-0.097	0.046	-0.178	7.732
CI_U	0.480	0.467	0.464	0.398	0.380	0.523	0.475	0.488	0.512	0.386	25.000

Table 2.26: Simulation results for the 3-state RSStC model with $T = 1,000$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the initial and transition probabilities.

	λ_1	λ_2	λ_3	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|3}$	$\pi_{2\|1}$	$\pi_{2\|2}$	$\pi_{2\|3}$	$\pi_{3\|1}$	$\pi_{3\|2}$	$\pi_{3\|3}$
True	0.333	0.333	0.333	0.700	0.200	0.100	0.200	0.700	0.100	0.100	0.200	0.700
Bias	0.021	0.027	-0.048	0.030	0.027	-0.057	-0.049	0.280	-0.230	-0.019	-0.029	0.048
RMSE	0.461	0.454	0.483	0.117	0.105	0.093	0.128	0.342	0.286	0.054	0.138	0.148
CI_L	0.000	0.000	0.000	0.348	0.026	0.034	0.059	0.002	0.052	0.026	0.032	0.319
CI_U	1.000	1.000	1.000	0.778	0.414	0.293	0.528	0.780	0.733	0.223	0.512	0.839

Table 2.27: Simulation results for the 3-state RSStC model with $T = 1,000$: true parameter value, bias, RMSE, lower and upper bounds of the CIs (CI_L and CI_U, respectively) of the dependence parameters, and of the number of degrees of freedom.

State 1	$\rho_1^{(12)}$	$\rho_1^{(13)}$	$\rho_1^{(14)}$	$\rho_1^{(15)}$	$\rho_1^{(23)}$	$\rho_1^{(24)}$	$\rho_1^{(25)}$	$\rho_1^{(34)}$	$\rho_1^{(35)}$	$\rho_1^{(45)}$	ν_1
True	0.900	0.700	0.800	0.800	0.750	0.900	0.800	0.700	0.800	0.800	3.000
Bias	0.008	0.008	0.007	0.010	0.008	0.008	0.010	0.008	0.009	0.009	-0.058
RMSE	0.032	0.027	0.036	0.053	0.025	0.028	0.038	0.025	0.036	0.046	0.767
CI_L	0.850	0.848	0.853	0.848	0.848	0.846	0.844	0.850	0.848	0.847	2.000
CI_U	0.936	0.932	0.933	0.933	0.934	0.935	0.935	0.933	0.933	0.934	4.886
State 2	$\rho_2^{(12)}$	$\rho_2^{(13)}$	$\rho_2^{(14)}$	$\rho_2^{(15)}$	$\rho_2^{(23)}$	$\rho_2^{(24)}$	$\rho_2^{(25)}$	$\rho_2^{(34)}$	$\rho_2^{(35)}$	$\rho_2^{(45)}$	ν_2
True	0.500	0.300	0.500	0.400	0.400	0.400	0.500	0.400	0.500	0.300	7.000
Bias	-0.084	-0.077	-0.074	-0.091	-0.088	-0.066	-0.082	-0.071	-0.078	-0.092	-0.879
RMSE	0.215	0.208	0.188	0.235	0.228	0.190	0.218	0.203	0.188	0.232	3.364
CI_L	0.141	0.136	0.220	0.045	0.065	0.199	0.135	0.155	0.243	0.037	2.858
CI_U	0.887	0.895	0.881	0.881	0.889	0.888	0.887	0.893	0.896	0.890	15.316
State 3	$\rho_3^{(12)}$	$\rho_3^{(13)}$	$\rho_3^{(14)}$	$\rho_3^{(15)}$	$\rho_3^{(23)}$	$\rho_3^{(24)}$	$\rho_3^{(25)}$	$\rho_3^{(34)}$	$\rho_3^{(35)}$	$\rho_3^{(45)}$	ν_1
True	0.100	0.150	0.050	0.050	-0.100	0.100	-0.050	0.050	0.100	-0.010	15.000
Bias	-0.038	-0.133	-0.102	-0.102	-0.102	-0.072	-0.041	-0.105	-0.071	-0.101	-1.924
RMSE	0.105	0.197	0.158	0.162	0.163	0.131	0.108	0.162	0.126	0.165	6.528
CI_L	0.024	-0.186	-0.062	-0.173	-0.172	0.036	0.037	-0.061	0.052	-0.193	7.927
CI_U	0.420	0.379	0.399	0.348	0.340	0.465	0.417	0.418	0.446	0.318	25.000

2.C Semi-parametric approach

As discussed in Section 2 of the present Chapter, a semi-parametric model presents a viable alternative to the parametric model when the primary objective is modeling the joint distribution rather than the marginal distributions. In the case of employing a semi-parametric model, normalized ranks are computed directly from the observed log-returns of each individual time-series.

In the following, we show the results of the RSStC model obtained adopting this alternative approach applied to cryptocurrencies log-returns. We select the number of hidden states based on the Integrated Complete Likelihood (ICL, Biernacki et al., 2000) criterion and, according to the values reported in Table 2.28 also showing the realized values of the Bayesian Information Criterion (BIC, Schwarz, 1978) we select an RSStC model with 2 hidden states. This result is coherent with that obtained in Section 5, as we select a 2-state model in both cases.

Table 2.28: Bayesian Information Criteria (BIC) and Integrated Completed Likelihood (ICL) computed for increasing values of the number of hidden regimes k (semi-parametric approach).

	k			
Information Criterion	1	2	3	4
ICL	-8,996.649	**-9,696.635**	-9,624.546	-9,494.576
BIC	-8,996.649	-9,876.457	-9,912.995	**-9,933.16**

Table 2.29 shows the estimated number of degrees of freedom and the determinant of the matrices of dependence parameters. Similarly to previous results, regimes with higher general correlation are associated with lower estimated values for ν_u.

Table 2.29: Estimated number of degrees of freedom ν_u, and determinant of the estimated matrices of dependence parameters under the 2-state RSStC model (semi-parametric approach). Standard errors (in brackets) are obtained with nonparametric block bootstrap.

State	$u = 1$	$u = 2$
ν_u	6.603 (1.808)	7.517 (3.763)
$\det(\boldsymbol{R}_u)$	0.001	0.089

Tables 2.30 and 2.31 report the estimated dependence parameters and Kendall's tau, respectively, under the 2-state RSS-tC model. We recover a correlation structure from these estimates almost identical to that provided by the 2-state RSStC model estimated in Section 5 (Tables 2.9 and 2.10).

Table 2.30: Estimated dependence parameters $\rho_u^{(ij)}$ under the 2-state RSStC model (semi-parametric approach). Standard errors (in brackets) are obtained with nonparametric block bootstrap.

State $u = 1$	BTC	ETH	XRP	LTC	BCH
BTC	1.000 (0.000)	-	-	-	-
ETH	0.905 (0.019)	1.000 (0.000)	-	-	-
XRP	0.859 (0.032)	0.895 (0.027)	1.000 (0.000)	-	-
LTC	0.892 (0.026)	0.912 (0.019)	0.903 (0.029)	1.000 (0.000)	-
BCH	0.887 (0.031)	0.903 (0.029)	0.893 (0.040)	0.923 (0.031)	1.000 (0.000)
State $u = 2$	BTC	ETH	XRP	LTC	BCH
BTC	1.000 (0.000)	-	-	-	-
ETH	0.634 (0.127)	1.000 (0.000)	-	-	-
XRP	0.441 (0.138)	0.545 (0.117)	1.000 (0.000)	-	-
LTC	0.655 (0.074)	0.704 (0.050)	0.531 (0.097)	1.000 (0.000)	-
BCH	0.548 (0.155)	0.620 (0.099)	0.452 (0.148)	0.637 (0.132)	1.000 (0.000)

Table 2.32 reports the estimated transition probability matrices for the 2-state RSStC model. Even in this case, strong persistence is observed: the highest off-diagonal element is given by the transition probability from state 2 to state 1 (0.091).

Table 2.33 shows the state-conditional means and standard deviations of the five cryptocurrencies log-returns with the state allocation obtained through the Viterbi (Viterbi, 1967) algorithm. The categorization of states as bearish and bullish aligns with the descriptions outlined in the Chapter. In fact, the first state is consistently characterized by negative log-returns, while the second state is marked by positive log-returns for all cryptocurrencies.

Regimes 1 and 2 are visited 65.67% and 34.33%, respectively, with sojourn times equal to 31 days for the first state and 20 days for the second state. These values are slightly different from the previous findings. In order to provide a more appropriate comparison of the time-series decoded with the two approaches, we also compute the adjusted Rand index (ARI, Hubert and Arabie, 1985) equal to 0.53: this result suggests that the choice made on the marginal distributions slightly affects the posterior allocation. For a more detailed explanation of how to compute this index, please refer to Appendix 4.B of Chapter 4.

Table 2.31: Kendall's tau computed with the estimated dependence parameters $\rho_u^{(ij)}$ under the 2-state RSStC model (semi-parametric approach).

State $u = 1$	BTC	ETH	XRP	LTC	BCH
BTC	1.000	-	-	-	-
ETH	0.720	1.000	-	-	-
XRP	0.658	0.706	1.000	-	-
LTC	0.701	0.731	0.718	1.000	-
BCH	0.695	0.717	0.703	0.748	1.000
State $u = 2$	BTC	ETH	XRP	LTC	BCH
BTC	1.000	-	-	-	-
ETH	0.437	1.000	-	-	-
XRP	0.291	0.367	1.000	-	-
LTC	0.455	0.497	0.356	1.000	-
BCH	0.369	0.426	0.298	0.439	1.000

Table 2.32: Estimated transition probabilities $\pi_{u|v}$ under the 2-state RSStC model (semi-parametric approach). Bootstrap standard errors are reported in brackets.

State	$u = 1$	$u = 2$
$v = 1$	0.949 (0.051)	0.051 (0.043)
$v = 2$	0.090 (0.051)	0.910 (0.043)

Table 2.33: State-conditional means and standard deviations of the five cryptocurrencies log-returns with state allocation obtained through the Viterbi algorithm.

State 1	Mean (%)	S.D. (%)	State 2	Mean (%)	S.D. (%)
BTC	-0.237	3.904	BTC	0.715	4.565
ETH	-0.396	5.105	ETH	1.010	5.461
XRP	-0.586	4.961	XRP	1.265	8.484
LTC	-0.576	5.182	LTC	1.107	6.334
BCH	-0.617	5.660	BCH	0.968	7.966

Bibliography

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

Ang, A. and Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63:443–494.

Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4:313–337.

Ardenas-Ovando, R., Noguez, J., and Rangel-Escareno, C. (2017). RcppHMM: Seamless R and C++ integration. *CRAN*.

Ardia, D., Bluteau, K., and Rüede, M. (2019). Regime changes in bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29:266–271.

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.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37:1554–1563.

Bauwens, L. and Laurent, S. (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business & Economic Statistics*, 23:346–354.

Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21:79–109.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

Breymann, W., Dias, A., and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3:1–14.

Cerqueti, R., Giacalone, M., and Mattera, R. (2020). Skewed non-Gaussian GARCH models for cryptocurrencies volatility modelling. *Information Sciences*, 527:1–26.

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, Newtown, PA.

Chollete, L., Heinen, A., and Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime-switching copula. *Journal of Financial Econometrics*, 7:437–480.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, pages 841–862.

Cremaschini, A., Punzo, A., Martellucci, E., and Maruotti, A. (2023). On stylized facts of cryptocurrencies returns and their relationship with other assets, with a focus on the impact of COVID-19. *Applied Economics*, 55:3675–3688.

Czado, C. and Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, 9:453–477.

Das, S. R. and Uppal, R. (2004). Systemic risk and international portfolio choice. *The Journal of Finance*, 59:2809–2834.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73:111–129.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–22.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20:134–144.

Du, Z. (2016). Nonparametric bootstrap tests for independence of generalized errors. *The Econometrics Journal*, 19:55–83.

Duffie, D. and Pan, J. (1997). An overview of Value-at-Risk. *Journal of derivatives*, 4:7–49.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18.

Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5:1–50.

Fischer, M., Köck, C., Schlüter, S., and Weigert, F. (2009). An empirical analysis of multivariate copula models. *Quantitative Finance*, 9:839–854.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42:2243–2281.

Garcia, R. and Ghysels, E. (1998). Structural change and asset pricing in emerging markets. *Journal of International Money and Finance*, 17:455–473.

Genest, C., Nešlehová, J., and Ziegel, J. (2011). Inference in multivariate Archimedean copula models. *Test*, 20:223–256.

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

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Härdle, W. K., Okhrin, O., and Wang, W. (2015). Hidden Markov structures for dynamic copulae. *Econometric Theory*, 31:981–1015.

Hernández, L., Tejero, J., and Vinuesa, J. (2014). Maximum likelihood estimation of the correlation parameters for elliptical copulas. *arXiv preprint*, arXiv:1412.6316:1–13.

Huang, J.-J., Lee, K. J., Liang, H., and Lin, W. F. (2009). Estimating value at risk of portfolio by conditional copula-GARCH method. *Insurance: Mathematics and Economics*, 45:315–324.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall, London.

Joe, H. (2014). *Dependence Modeling with Copulas*. CRC press, Boca Raton, FL.

Joe, H. and Kurowicka, D. (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapore.

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. *University of British Columbia, Department of Statistics, Technical Report*, 166:1–22.

Jondeau, E. and Rockinger, M. (2006). The copula-GARCH model of conditional dependencies: An international stock market application. *Journal of International Money and Finance*, 25:827–853.

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

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93.

Koki, C., Leonardos, S., and Piliouras, G. (2022). Exploring the predictability of cryptocurrencies via Bayesian hidden Markov models. *Research in International Business and Finance*, 59:101554.

Kristoufek, L. (2013). Bitcoin meets Google trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3:1–7.

Longin, F. and Solnik, B. (2002). Extreme correlation of international equity markets. *The Journal of Finance*, 56:649–676.

Maruotti, A. and Punzo, A. (2021). Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review*, 89:447–480.

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

Nasri, B. R. and Rémillard, B. N. (2019). Copula-based dynamic models for multivariate time series. *Journal of Multivariate Analysis*, 172:107–121.

Nasri, B. R., Rémillard, B. N., and Thioub, M. Y. (2020). Goodness-of-fit for regime-switching copula models with application to option pricing. *The Canadian Journal of Statistics*, 48:79–96.

Nystrup, P., Lindström, E., and Madsen, H. (2020). Learning hidden Markov models with persistent states by penalizing jumps. *Expert Systems with Applications*, 150:113307.

Okimoto, T. (2008). New evidence of asymmetric dependence structures in international equity markets. *Journal of Financial and Quantitative Analysis*, 43:787–816.

Ötting, M., Langrock, R., and Maruotti, A. (2021). A copula-based multivariate hidden Markov model for modelling momentum in football. *AStA Advances in Statistical Analysis*, pages 1–19.

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, 2:130–168.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Pennoni, F., Bartolucci, F., Forte, G., and Ametrano, F. (2021). Exploring the dependencies among main cryptocurrency log-returns: A hidden Markov model. *Economic Notes*, 51:e12193.

Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22:270–293.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rémillard, B. (2011). Validity of the parametric bootstrap for goodness-of-fit testing in dynamic models. *Available at SSRN 1966476*, pages 1–43.

Remillard, B. (2013). *Statistical Methods for Financial Engineering*. CRC press, Boca Raton, FL.

Rodriguez, J. C. (2007). Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, 14:401–423.

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

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

Simard, C. and Rémillard, B. (2015). Forecasting time series with multivariate copulas. *Dependence Modeling*, 3:59–82.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231.

Stöber, J. and Czado, C. (2014). Regime switches in the dependence structure of multidimensional financial data. *Computational Statistics & Data Analysis*, 76:672–686.

Telli, Ş. and Chen, H. (2020). Structural breaks and trend awareness-based interaction in crypto markets. *Physica A: Statistical Mechanics and its Applications*, 558:124913.

Theodossiou, P. (2015). Skewed generalized error distribution of financial assets and option pricing. *Multinational Finance Journal*, 19:223–266.

Timmermann, A. (2018). Forecasting methods in finance. *Annual Review of Financial Economics*, 10:449–479.

Tiwari, A. K., Aye, G. C., Gupta, R., and Gkillas, K. (2020). Gold-oil dependence dynamics and the role of geopolitical risks: Evidence from a Markov-switching time-varying copula model. *Energy Economics*, 88:104748.

Trede, M. (2020). Maximum likelihood estimation of high-dimensional Student-t copulas. *Statistics & Probability Letters*, 159:108678.

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

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

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R.* CRC press, Boca Raton, FL.

Chapter 3

What drives cryptocurrency returns? A sparse statistical jump model approach[1]

3.1 Introduction

Since the introduction of Bitcoin in 2009, which operates with a decentralized ledger system known as the blockchain, cryptocurrencies have attracted much attention, and today some market participants argue that they constitute a separate asset class (Bianchi, 2020; Pele et al., 2021). Following the introduction of Bitcoin, many other cryptocurrencies have been implemented, collectively referred to as altcoins. By March 2022, there were approximately 18K cryptocurrencies in total.

Time-series analysis of cryptocurrency dynamics show they exhibit regime switching, structural breaks and jumps in both returns and volatility (Ardia et al., 2019; Chaim and Laurini, 2018; Shen et al., 2020) (see Figure 3.1). In his seminal work, Hamilton (1989) suggests that the dynamics of financial returns can be described by Markovian regime-switching processes, with drastic breaks associated with events like economic crises or political events. Later, Rydén et al. (1998) demonstrate empirically that a simple *hidden Markov model* (HMM) can reproduce most of the common stylized facts in asset returns (cf. Cont (2001) and Lindström et al. (2015, Chapter 1)).

Figà-Talamanca et al. (2021) adopt a multivariate approach to demonstrate the presence of common market regimes amongst cryptocurrencies. They analyze first differences of Bitcoin, Ethereum, Litecoin, and Monero prices, and conclude that a multivariate generalized white noise Markov switching model with three states best fits the data. Moreover, they characterize the regimes in terms of their different state-conditional volatilities. Koki et al. (2022) estimate several different HMMs for the purpose of forecasting Bitcoin, Ethereum and Ripple and demonstrate that a four-state specification provides the best out-of-sample

[1]This Chapter has been published in: Cortese F., Kolm P., Lindström E. (2023). What drives cryptocurrency returns? A sparse statistical jump model approach. *Digital Finance*, 1-36. https://link.springer.com/article/10.1007/s42521-023-00085-x

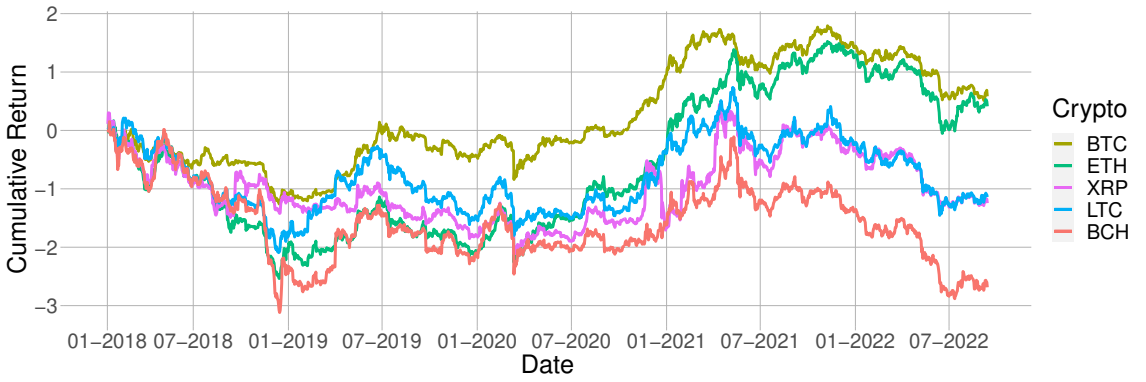


Figure 3.1: Cumulative log-returns of the five cryptocurrencies Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Bitcoin Cash (BCH), over the period January 2018 - September 2022.

performance. However, the statistical properties of the hidden states are not consistent across the three cryptocurrencies, making an interpretation of the latent states as different economic regimes difficult. An alternative explanation to their empirical results is that the switching dynamics is governed by some exogenous process(es) not included in their models.

While the understanding of the key drivers of cryptocurrency returns is still emerging, the literature has identified a number of crypto-, sentiment- and financial market-related features that impact cryptocurrency market dynamics (see, for example Koki et al. (2022); Yae and Tian (2022) for an overview). Kristoufek (2015) finds that trade and transaction volumes are positively correlated with Bitcoin returns. Similarly, Aalborg et al. (2019) suggest that frequent use of Bitcoin would increase its demand, but also that this relationship might be negative during periods of high volatility. Catania et al. (2019) and Bianchi (2020) study the dependence of many cryptocurrencies on macroeconomic factors like S&P500 returns, market volatility (VIX) and commodities. They find a weak correlation between returns on cryptocurrencies and commodities, especially gold. Yae and Tian (2022) demonstrate that the change in correlation between Bitcoin and S&P500 returns predicts future Bitcoin returns. In particular, they show that an increase (decrease) in correlations today suppresses (boosts) Bitcoin prices the following day. Xiong et al. (2020) provide evidence of the importance of including production cost and use value (a measure of how many people use a particular coin as a mean of exchanging value) of Bitcoin, as measured by the number of unique addresses and the Bitcoin hash rate. Several studies suggest that market sentiment plays a crucial role in determining the prices of cryptocurrencies. Kristoufek (2013) finds a strong correlation between the number of visits to the Bitcoin Wikipedia page and price dynamics. Similarly, Aalborg et al. (2019) and Cheah et al. (2020) suggest that in-

vestor attention, as measured by Google search queries for the term "bitcoin", also impacts the cryptocurrency market. Urquhart (2018) analyzes Google Trends search queries and suggests that large realized volatility and trading volumes increase public attention towards BTC. Likewise, Figa-Talamanca and Patacca (2019) observe that the search volume index provided by Google is a statistically significant predictor of the conditional variance of BTC returns.

Liu and Tsyvinski (2021) and Liu et al. (2022) find that cross-sectional factors constructed from market return, size, momentum, and public attention can forecast cryptocurrency returns. Cheah et al. (2020) and Koki et al. (2022) also find that time-series momentum is a statistically significant predictor for the upward and downward trending states of crypto markets.

In this Chapter, we aim to determine what are the most important drivers of the return dynamics of the five largest and most liquid cryptocurrencies; namely, Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Bitcoin Cash (BCH). For this purpose, we construct a large set of candidate features grouped into three categories: financial market, sentiment and crypto market-related features. While many of these features are drawn from the emerging literature, we also propose some that are new.

We adopt the *sparse statistical jump model* (SJM) of Nystrup et al. (2021) to select the features that best explain cryptocurrency returns. As an alternative modeling framework to HMMs, the SJMs have several advantages. First, they allow us to *simultaneously* perform parameter estimation, state-sequence decoding and feature selection. Second, Nystrup et al. (2021) show that, in contrast to many "competing" models, SJMs are more robust to misspecification and initialization, deliver acceptable performance even on smaller samples, and tend to be more efficient for high-dimensional feature vectors. Third, the estimated SJMs are easy to interpret from their conditional state dynamics and the weight associated with each feature, providing an opportunity to reconcile statistical properties and selected features with observed market behavior and economic intuition.

We demonstrate in our empirical study that when more than four hundred features are included, a three-state SJM best explains cryptocurrency returns. The three states have intuitive economic interpretations, corresponding to bull, neutral, and bear market regimes. The relevant features selected by the SJM are exponential moving averages of returns, features representing trend and reversal signals drawn from the technical analysis literature, market activity and public attention. The features that we use for representing market activity and public attention are new. Our findings are consistent with those of Figà-Talamanca et al. (2021), who also identify three common market regimes. Comparing our decoded state sequences with theirs, we find that they are similar. However, in contrast to their results, we do not find the volatilities to be the key features that distinguish the

market states.

The rest of the Chapter is organized as follows. In Section 3.2, we review the mathematical formulation of SJMs, their implementation and hyperparameter tuning. In Section 3.3, we describe the candidate features we use in inferring drivers of cryptocurrency returns. We describe the preparation of our dataset and present the results from our empirical analysis in Section 3.4. Section 3.5 concludes. Appendix 3.A provides a complete list of all the features we use in this study.

3.2 Methodology

Traditional regime switching models, such as HMMs that have been used for decades in finance and economics, strike a balance between being interpretable and flexible enough to model complex non-stationary behavior. Well-known limitations of this class of models include sensitivity to model misspecification and feature selection for exogenous variables (see Zucchini et al. (2017) for a recent overview). Another issue is the difficulty to reliably estimate model parameters, as the log-likelihood function is notoriously multimodal. Estimation frameworks, such as direct maximization of the log-likelihood, the expectation-maximization algorithm, and Bayesian approaches are compared in Rydén (2008), with no uniform preference of any particular framework over the others.

Recently, Bemporad et al. (2018) introduced the class of *statistical jump models* (JM), nesting e.g. HMMs. We remark that JMs are not related to jump-diffusion models, a common class of stochastic processes. Following the trend in statistics and machine learning of reformulating probabilistic models as well-behaved optimization problems, such as LASSO or support vector machines, a JM is estimated by minimizing the combined loss of the (negative) likelihood and penalties for jumping between states.

Nystrup et al. (2020b) reformulate the JM as a temporal clustering model. They show in a simulation study that the JM interpretation of an HMM has many advantages, including robustness against model misspecification and poor initialization of the optimizer, as well as surprisingly rapid convergence, typically within only a few iterations. Nystrup et al. (2020b) proves that the global loss can be optimized by sequentially alternating between (a) fitting the model parameters while keeping the state sequence fixed, and (b) estimating the state sequence through dynamic programming while keeping the model parameters fixed.

In a further extension, Nystrup et al. (2021) introduce the *sparse statistical jump model* (SJM) by incorporating feature selection from the clustering literature. This is important as large scale feature selection has been infeasible for standard HMMs. They demonstrate in a series of simulation studies that SJMs outperform competing frameworks and are capable of recovering true features, while rejecting false features with high probability even when considering hundreds of them.

3.2.1 Mathematical formulation of the SJM

The JM proposed by Bemporad et al. (2018) is governed by a latent state sequence, $\{s_t\}$, switching between K states, each associated with a vector of parameters, $\{\theta_k\}_{k=1}^K$. The model is fitted to a time-series $\boldsymbol{y}_1, \dots, \boldsymbol{y}_T$, by minimizing the loss

$$\sum_{t=1}^{T-1} \left[l(\boldsymbol{y}_t; \theta_{s_t}) + \lambda \mathbb{I}_{\{s_t \neq s_{t-1}\}} \right] + l(\boldsymbol{y}_T; \theta_{s_T}) \,, \tag{3.1}$$

where the hyperparameter $\lambda \geq 0$ controls the number of jumps between states and $l(\cdot)$ is some loss function to be specified. Bemporad et al. (2018) show in their paper that a suitable choice of these local likelihood terms can generate a number of well known model classes, including HMMs.

Nystrup et al. (2020b) consider temporal clustering by transforming the data into p standardized features, $\widetilde{\boldsymbol{y}}_{t,p} \in \mathbb{R}^p$, and using a quadratic loss function. Their corresponding objective function is given by

$$\sum_{t=1}^{T-1} \left[\|\widetilde{\boldsymbol{y}}_{t,p} - \boldsymbol{\mu}_{s_t}\|_2^2 + \lambda \mathbb{I}_{\{s_t \neq s_{t-1}\}} \right] + \|\widetilde{\boldsymbol{y}}_{T,p} - \boldsymbol{\mu}_{s_T}\|_2^2 \,, \tag{3.2}$$

where $\boldsymbol{\mu}_{s_t}$ is the conditional mean of state s_t. It can be shown that this form of temporal clustering collapses into K-means clustering when $\lambda = 0$.

Nystrup et al. (2021) note that the *total sum of squares* (TSS) can be expressed as the sum of *within-cluster sum of squares* (WCSS) and the *between-cluster sum of squares* (BCSS)

$$\sum_{t=1}^{T} \|\widetilde{\boldsymbol{y}}_{t,p} - \bar{\boldsymbol{\mu}}\|_2^2 = \sum_{t=1}^{T} \|\widetilde{\boldsymbol{y}}_{t,p} - \boldsymbol{\mu}_{s_t}\|_2^2 + \sum_{k=1}^{K} n_k \|\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}\|_2^2 \,, \tag{3.3}$$

where $\bar{\boldsymbol{\mu}}$ is the unconditional mean of the features, and $\boldsymbol{\mu}_k$ the conditional mean of the features in the k-th state.

Building upon the work by Witten and Tibshirani (2010) on feature selection in clustering, and using that minimizing the WCSS in Equation (3.2) is equivalent to maximizing the BCSS (as the TSS is constant), Nystrup et al. (2021) propose to solve

$$\begin{aligned} \max_{\boldsymbol{\mu}_k, \{s_t\}, \boldsymbol{w}} \quad & \boldsymbol{w}' \textstyle\sum_{k=1}^{K} n_k \left(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}\right)^2 - \lambda \sum_{t=1}^{T-1} \mathbb{I}_{\{s_t \neq s_{t-1}\}} \\ \text{subject to} \quad & \|\boldsymbol{w}\|^2 \leq 1, \quad \|\boldsymbol{w}\|_1 \leq \kappa \\ & w_p \geq 0 \quad \forall p \,, \end{aligned} \tag{3.4}$$

where $\boldsymbol{w}$ is the vector of feature weights, n_k denotes the number of observations belonging to the k-th cluster, and the hyperparameter $1 \leq \kappa \leq \sqrt{p}$ controls the degree of sparsity

of the features. Nystrup et al. (2021) show that this optimization problem can be solved by iteratively alternating between (a) fitting model parameters given the state sequence, $\{s_t\}$, and weights $\boldsymbol{w}$, (b) deriving the state sequence, $\{s_t\}$, by solving a dynamic program, essentially running the Viterbi algorithm backwards, and (c) updating the weights $\boldsymbol{w}$ using soft thresholding.

3.2.2 Model implementation and hyperparameters

As the SJM is unsupervised, we select its hyperparameters based on a version of the *generalized information criterion* (GIC) (Fan and Tang, 2013) for high-dimensional, penalized models suitably modified for SJMs.

Fan and Tang (2013) consider a setup where the number of features is considerably larger than the number of observations. They define a GIC as

$$\text{GIC} = \frac{1}{T}\left\{2\left(\ell^S(\boldsymbol{Y}) - \ell(\hat{\theta}_{\boldsymbol{\alpha}}, \boldsymbol{Y})\right) + a_T M\right\}, \tag{3.5}$$

where $\boldsymbol{Y}$ is the matrix of features, M is a measure of model complexity, a_T is a penalty term possibly depending on the number of observations T and features p, $\ell(\hat{\theta}_{\boldsymbol{\alpha}}, \boldsymbol{Y})$ is the log-likelihood computed using the estimated active set of features, $\boldsymbol{\alpha}$, and $\ell^S(\boldsymbol{Y})$ is the log-likelihood for the saturated model, which is the model obtained considering the entire set of features. Fan and Tang (2013) note that the generalized versions of the *Akaike's information criterion* (AIC) (Akaike, 1974) or the *Bayesian information criterion* (BIC) (Schwarz, 1978) are recovered by setting M equal to the total number of parameters, and $a_T = 2$ or $a_T = \log(T)$ for AIC and BIC, respectively.

In this Chapter, we determine the optimal model using our preliminary findings on a GIC customized for SJMs. Notably, we employ $a_T = \log(T)$, transforming it into a modified variant of the BIC. However, in Chapter 4, we introduce a more rigorous version of the GIC, employing a slightly different expression. For a comprehensive understanding of the preliminary GIC version and the results of our simulation study, which evaluated its effectiveness in selecting the correct model, please refer to Cortese et al. (2023).

The modified BIC (which we denote by BIC in the following) is defined as

$$\text{BIC} = \frac{1}{T}\left\{2\left(L_T(\bar{\lambda}, \bar{\kappa}, \bar{K}; \boldsymbol{Y}) - L_T(\lambda, \kappa, K; \boldsymbol{Y})\right) + M\log(T)\right\}, \tag{3.6}$$

where $L_T(\lambda, \kappa, K; \boldsymbol{Y})$ denotes the estimated BCSS, λ, κ, and K are the values for the jump penalty, sparsity hyperparameter, and number of latent states, respectively. The terms $\bar{\lambda}$, $\bar{\kappa}$, and $\bar{K}$ represent the SJM hyperparameters for the saturated model, which includes all features. Setting $\bar{\lambda} = 0$ and $\bar{\kappa} = \sqrt{p}$ is straightforward, but choosing $\bar{K}$ is less clear. In our experience, for modeling recurrent states, we recommend not exceeding $\bar{K} = 6$. This is

because when $\bar{K}$ is too high, the index tends to select numerous states, each visited only once. We define M in Equation (3.6) by

$$M = K_0|\boldsymbol{\alpha}_0| + |\boldsymbol{\alpha}_0|(K - K_0) + K_0(|\boldsymbol{\alpha}_\kappa| - |\boldsymbol{\alpha}_0|) + \sum_t \mathbb{I}_{s_t \neq s_{t-1}}\,, \tag{3.7}$$

where the three first terms come from a linear approximation of K and $|\boldsymbol{\alpha}_\kappa|$ near the point $(K_0, |\boldsymbol{\alpha}_0|)$. This expression penalizes for increasing values of K and $|\boldsymbol{\alpha}_\kappa|$, the number of latent states and active features. $|\boldsymbol{\alpha}_\kappa|$ indirectly depends on the hyperparameter κ, and it increases with increasing values of κ. The last term in Equation (3.7) counts the number of jumps across states and thereby depends indirectly on the jump penalty λ.

In practical applications, we suggest to select K_0 and $|\boldsymbol{\alpha}_0|$ based on prior knowledge of the number of latent states and relevant features. In our empirical work, we set $|\boldsymbol{\alpha}_0| = 150$ as we surmise that first and second order moments, correlations, volumes and momentum related features may be relevant, and these features are roughly 30 for each of the five cryptocurrencies. The choice of K_0 is based primarily on qualitative properties. Selecting it too small will restrict the dynamics. Too large and the model does not generate persistent or even recurrent states. Rather, it segments the time-series into separate blocks, thus negatively impacting model predictability. Koki et al. (2022) find that a non-homogeneous HMM with four states best describes BTC, ETH and XRP log-returns dynamics. In his comparative analysis, Bulla (2011) observes that considering an HMM with Student-t conditional distributions results into selecting a fewer number of regimes. In fact, compared to Gaussian conditional distributions, model estimation is less dependent on a few extreme observations that might cause the number of states to increase. Nystrup et al. (2020b) show that the JM is robust against distributional misspecification, similar to using an HMM with Student-t conditional distribution. Hence, we use $K_0 = 3$, reflecting our prior assumption of a modest number of states. In our setting, assuming an *a priori* number of states equal to 2, 3 or 4 does not significantly change the results.

3.3 Econometric features

We construct a large set of features as potential candidates for explaining cryptocurrency returns. Many of these features have been proposed in the literature (see, Yae and Tian (2022) for a survey), but some are new. We categorize them into three groups: financial market, sentiment, and crypto market-related features. Financial market features are based on information from the equity, fixed income, foreign exchange and commodity markets. Sentiment features proxy for investor attention toward the cryptocurrency markets. Crypto market-related features include prices and volumes of the cryptocurrencies as well as metrics

related to the blockchain.

Some of our features require parameter choices before they are calculated, such as window length and the number of observations. This is addressed in our empirical study by including multiple versions of the same variable computed for different parameters, from which the feature selection algorithm then can choose the most suitable appropriate features. Next, we describe the construction of the features in each group.

3.3.1 Financial market features

There is an ongoing debate whether cryptocurrencies constitute a separate asset class, with some market participants arguing that due fat-tails, high kurtosis and conditional volatility of their returns (Pele et al., 2021), they behave significantly differently than traditional assets. Nevertheless, it is natural to ask whether there exists some relationship between cryptocurrencies and traditional assets classes. Therefore, we construct a number of features based on time-series from the equity, fixed income, foreign exchange and commodity markets, aimed at describing cross-asset dynamics between cryptocurrencies and traditional markets. In particular, we consider the following features: first differences of WTI oil prices; first differences of 10-year minus 3-months constant maturity Treasury yields (T10Y3M); log-returns of gold; log-returns of the S&P500; log-returns of NASDAQ; log-returns of EUR-USD; log-returns of JPY-USD; log-returns of CNY-USD; log-differences of VIX index; and *exponential moving averages* (EMAs) for log-differences of VIX index with half-lives $d =$ 1, 2, 7, and 14 days. To proxy for possible comovements of cryptocurrencies relative to traditional asset classes (Selmi et al., 2018), we include *exponentially weighted linear and Gerber correlations*, denoted by ρ_d and g_d, (Gerber et al., 2022) of BTC log-returns with all other financial market features above with half-lives $d = 1$, 2, 7, and 14 days.

In a distinct analysis, in response to the reviewer's suggestion, we incorporate one additional variable related to inflation expectations, such as the 10-year breakeven inflation rate. However, we choose not to include the 5-year forward inflation expectation rate due to its high linear coefficient correlation of 0.92 with the initial variable, implying that one could serve as a proxy for the other. Notably, the results remain consistent even after introducing this variable into the analysis.

3.3.2 Sentiment features

Cheah et al. (2020) suggest that investor sentiment, such as public attention, has an impact on the cryptocurrency markets. Likewise, Aalborg et al. (2019) find that Google searches for BTC can predict BTC trading volume, while Urquhart (2018) and Figa-Talamanca and Patacca (2019) provide evidence of a relationship between Google searches and volatility of BTC returns.

To proxy for public attention toward cryptocurrencies, we use the log-differences of the Google Trends indexes (GT) from the queries "bitcoin", "ethereum", "ripple", "litecoin", and "bitcoin cash". We also add a second set of public attention-related features computed as exponentially weighted linear and Gerber correlations with half-lives $d = 1$, 2, 7, and 14 days of log-differences of Google Trend indexes and log-returns of each cryptocurrency.

In response to the reviewer's guidance, we conduct an additional analysis incorporating the infectious disease equity market volatility tracker (Baker and Bloom, 2013). However, we refrain from including the economic policy uncertainty (EPU) indicator due to its reliance on monthly observations, whereas our data operates on a daily frequency. Notably, the empirical findings remain unchanged even with the inclusion of this variable.

3.3.3 Crypto market-related features

This category covers features derived directly from cryptocurrency prices, trade volumes, and blockchain-related metrics. Log-differences of the total number of unique addresses with balance (AddWB) used on the blockchain aim at measuring the use value of a given coin. The number of addresses with balance is defined as the number of unique identifiers that serves as a virtual location where the coin can be sent. This metric differs from the total number of unique addresses in that it only counts wallets currently holding a particular coin, while the other one considers all addresses ever created. We use first differences of the total volume on chain (VOC), as a feature for the aggregate volume of transactions recorded on chain. Following Cong et al. (2021), we construct three value factors as the ratio between the total number of addresses and prices (AM), the number of addresses with balance and prices (UM), and the recorded volume on chain and prices (TM). We compute all the above mentioned features only for BTC, ETH, LTC and BCH due to data availability issues.

Hash rates refer to the amount of computing power used by the cryptocurrency network to process transactions and serves as a measure of the production cost of the mining process (Xiong et al., 2020). We include log-differences of hash rates (HR) for BTC and ETH.

To capture first and second moments of cryptocurrency returns, we include USD denominated daily log-returns, and EMAs of log-returns and volatilities with half-lives $d =$ 1, 2, 7 and 14 days. To represent market activity, we use first differences of the logarithm of USD denominated trading volumes (V) and the corresponding EMAs with the same half-lives as above.

The results from several studies suggest that time-series momentum is an important driver of crypto-returns (see, for example Liu and Tsyvinski (2021); Liu et al. (2022); Yae and Tian (2022)). Therefore, we include several different momentum-based features in this study. Specifically, for each of the cryptocurrencies we consider the time-series

momentum signal (RF) of Moskowitz et al. (2012) which is based on time-series regressions with a variable lag of l. In our empirical, work we use $l = 1, 2, 7$ and 14 days. Moreover, taking inspiration from the technical analysis literature, we include the relative strength index (RSI) and the moving average converge-divergence minus signal (MACDS) indicator (Wilder, 1978; Appel, 2005). The MACDS and RSI are features that represent trend and reversal signals, respectively, and are often used together to determine whether markets are in either a trending or range-bound condition. We apply the standard parameter choices when computing these features.

To proxy for illiquidity, we include the Amihud (2002) illiquidity measure (AMIHUD), computed as the ratio of absolute daily log-returns and daily volumes for each coin.

Finally, we also construct exponentially weighted linear and Gerber correlations with half-lives $d = 1$, 2, 7 and 14 days of (a) BTC log-returns and log-returns of all other cryptocurrencies to obtain estimates of market betas, (b) log-returns and log-differences of trade volumes for each crypto, (c) BTC and ETH log-returns and log-differences of their corresponding hash rates, (d) BTC, ETH, LTC and BCH log-returns and their corresponding VOC and AddWB.

3.4 Empirical study

3.4.1 Data

The study by Alexander and Dakos (2020) emphasizes that, for empirical analysis of cryptocurrency markets, the choice of data sources is critical. In particular, they advise that researchers use trade data obtained from the crypto exchanges, rather than non-trade data from coin-ranking websites and other sources where data quality is significantly lower. Barucci et al. (2022) highlight that for intraday settings, cryptocurrencies quoted against BTC, ETH or Tether (USDT) are more liquid and therefore tend to be more accurate than those quoted against the dollar. In fact, USDT facilitates trades in cryptocurrencies as fees are lower and no bank transfers are needed.

In the present work, we take the perspective of a USD denominated investor and therefore use USD prices and volumes.[2] Following Pennoni et al. (2021), we use price and trade volume data from the Crypto Asset Lab (CAL)[3], who collect data from crypto exchanges

[2]To ensure that the results of our study is not an artefact of possible differences in daily USD vs. USDT cryptocurrency quotations, we also perform our analysis with USDT denominated daily prices and volumes obtained from KuCoin. Due to data availability issues, the data for this comparative analysis covers a shorter time period, from March 7, 2019 through September 13, 2022. We observe no significant differences between the model estimated with USD or USDT denominated data. The results of this comparative analysis are available upon request.

[3]The Crypto Asset Lab is an independent lab established at the University of Milano-Bicocca; see https://cryptoassetlab.diseade.unimib.it

that satisfy their reliability and liquidity criteria. We are grateful to CAL for providing us with daily volume-weighted USD denominated prices and aggregate trade volumes, recorded at midnight UTC, from the Coinbase-pro, Poloniex, Bitstamp, Gemini, Bittrex, Kraken, and Bitflyer digital exchanges.

We obtain treasury constant maturity yields data from FRED[4]; gold and WTI oil prices, S&P500, NASDAQ and VIX levels, EUR-USD, JPY-USD and CNY-USD exchange rates from Bloomberg; hash rates, number of unique addresses with balance and volumes on chain from intotheblock.com; and online search trends for the five cryptocurrencies from Google Trends[5].

Our dataset spans the time period from January 16, 2018 through September 13, 2022, with a total number of 1,702 daily observations. As traditional financial markets are only open during business days, their corresponding time-series lack values during weekends. For convenience, we impute any missing data using the `mice` **R** package (Van Buuren and Groothuis-Oudshoorn, 2011). Tables 3.1 and 3.2 provide the unconditional means, standard deviations and correlation matrix of the five cryptocurrency log-returns. From the different time-series, we derive a total of 409 features, all of which are stationary (see Section 3.3 for a description of the construction of the features and Appendix 3.A for a complete list).

Table 3.1: Daily unconditional means and standard deviations (SD) of the five cryptocurrency log-returns.

	Mean (%)	SD (%)
BTC	0.02	4.05
ETH	0.01	5.25
XRP	-0.09	6.03
LTC	-0.08	5.48
BCH	-0.18	6.24

3.4.2 Results

We use the `Python` implementation of the SJM from Nystrup et al. (2021), available from their online supplementary material.[6] The estimation of the SJM requires 3.1 seconds on a 16-core Intel i7-8750H with 16GB of RAM.[7] For the remainder of this section, we discuss the results from the model.

[4] https://fred.stlouisfed.org

[5] https://trends.google.com

[6] https://www.sciencedirect.com/science/article/pii/S0957417421009647#appSB

[7] We thank the University of Milano-Bicocca Data Science Lab (datalab) for supporting this work by providing computational resources.

Table 3.2: Unconditional correlation matrix of the five cryptocurrency log-returns.

	BTC	ETH	XRP	LTC	BCH
BTC	1.00	-	-	-	-
ETH	0.84	1.00	-	-	-
XRP	0.65	0.70	1.00	-	-
LTC	0.81	0.84	0.71	1.00	-
BCH	0.78	0.81	0.68	0.83	1.00

Based on the BIC as in Equation (3.6), we select the model with $K = 3$ states and with $\lambda = 5$, $\kappa = 4$. The daily state-conditional means and standard deviations of the returns are noticeably increasing from state 1 to state 3, as shown in Table 3.3. In Table 3.4, we observe

Table 3.3: State-conditional means and standard deviations (SD) of the five cryptocurrency log-returns obtained from the SJM model.

	State 1		**State 2**		**State 3**	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
BTC	1.02	3.64	-0.26	3.15	-1.39	5.72
ETH	1.36	4.55	-0.29	4.23	-2.05	7.38
XRP	0.87	6.27	-0.26	4.80	-1.69	7.42
LTC	1.22	5.22	-0.47	4.17	-1.87	7.51
BCH	1.31	6.15	-0.46	4.59	-2.57	8.37

that the correlations of the cryptocurreny log-returns increase from the first through the third state, where the correlations in the third state are remarkably high. This increase in correlations is consistent with the asymmetric correlation phenomena *within* equities and other asset classes, as well as *across* asset classes (see, for example Erb et al. (1994); De Bandt and Hartmann (2000); Cappiello et al. (2006)).

Figure 3.2 depicts the decoded state sequence together with the cumulative log-returns of BTC, ETH, XRP, LTC and BCH. Throughout our sample of $1,702$ observations, the model spends $38.9\%, 41.8\%$ and 19.3% of its time in each of the three different states, where each visit lasts for an average of $26.5, 17.8$ and 16.4 days.

Together, the conditional statistics above suggest an interpretation of the states as distinct market regimes, where the first, second and third states represent a bull, neutral, and bear market regime, respectively. While the first regime (bull market) has positive average return and moderate volatility, the second regime (neutral market) is characterized by an average return slightly below zero under moderate volatility but with higher

Table 3.4: State-conditional correlations of the five cryptocurrency log-returns obtained from the SJM model.

State 1	BTC	ETH	XRP	LTC	BCH
BTC	1.00	-	-	-	-
ETH	0.72	1.00	-	-	-
XRP	0.50	0.57	1.00	-	-
LTC	0.70	0.73	0.58	1.00	-
BCH	0.63	0.67	0.54	0.73	1.00
State 2	BTC	ETH	XRP	LTC	BCH
BTC	1.00	-	-	-	-
ETH	0.84	1.000	-	-	-
XRP	0.62	0.66	1.00	-	-
LTC	0.83	0.86	0.67	1.00	-
BCH	0.81	0.81	0.65	0.84	1.00
State 3	BTC	ETH	XRP	LTC	BCH
BTC	1.00	-	-	-	-
ETH	0.92	1.00	-	-	-
XRP	0.81	0.87	1.00	-	-
LTC	0.89	0.92	0.89	1.00	-
BCH	0.89	0.90	0.86	0.91	1.00

correlations than the first regime. In contrast, the third regime (bear market) is associated with a significant average negative return and high volatility, approximately twice the magnitude of the volatilities observed in the two other regimes. In addition, cryptocurrency returns are highly correlated in the bear market regime, suggesting that there is little to no cross-sectional diversification during bad times.

Next, we turn to examining the features selected by the SJM. From the 409 features, the model selects 19 as being relevant (see, Figure 3.3 for a depiction of the feature weights of the relevant features). We observe that the 7- and 14-day EMAs of log-returns are the most relevant, with RSIs following thereafter. The state-conditional values of the relevant features are given in Table 3.5. These values are consistent with the bull (positive trend), neutral (range-bound) and bear (negative trend) regime interpretations above. The state-conditional 7- and 14-day EMAs for log-returns are similar in sign and magnitude to the state-conditional means and are consistent with the upward and downward momentum observed in the first and third regime.

The state-conditional RSI for BTC are 66.46, 45.03 and 32.58, consistent with the bull (positive trend), neutral (range-bound) and bear (negative trend) regime interpretations

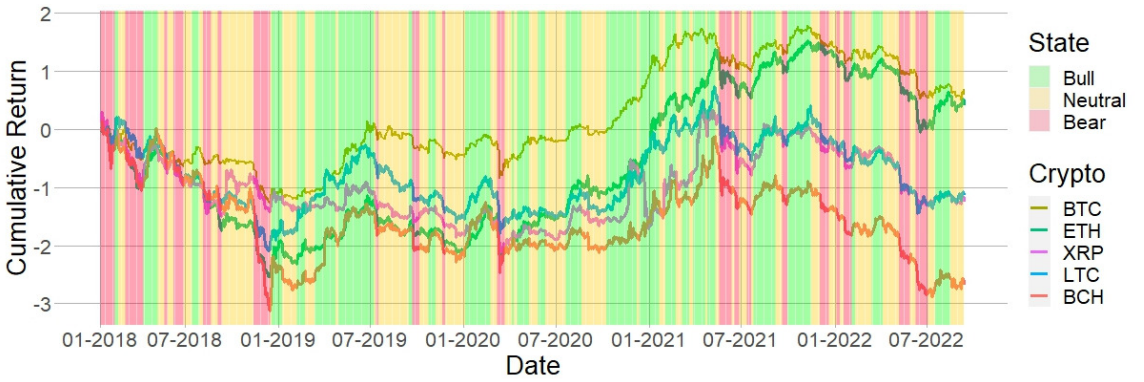


Figure 3.2: Cumulative log-returns of BTC, ETH, XRP, LTC, and BCH over the period January 2018 - September 2022, together with the state sequence obtained from the SJM in green (bull), yellow (neutral) and red (bear).

above. State-conditional RSIs for the other cryptocurrencies have similar magnitudes.

We observe that the 7-day exponentially weighted correlation of log-differences of Google Trend index and log-returns of BTC is relevant. The state-conditional values are equal to $0.26, -0.10$ and -0.38 in the bull, neutral and bear market states, respectively. This suggests that public attention affects the evolution of crypto markets, especially during downward and upward market trends. The model selects the 7- and 14-day exponentially weighted correlation of log-differences of Google Trend index and log-returns of ETH, having state-conditional values similar to that of $\rho_7(\mathrm{GT}_{\mathrm{BTC}}, r_{\mathrm{BTC}})$. Similarly, the 7- and 14-day exponentially weighted linear correlations of log-returns and log-differences of the BTC and ETH trade volumes are also selected. The state-conditional values in the bull, neutral and bear market states of the 7-day correlations for BTC are 0.24, -0.07 and -0.36, respectively. The corresponding values for the 14-day correlations of BTC and for the 7- and 14-day correlations of ETH are similar in magnitude. Finally, we note the model does not select any features from the group of financial market features (cf. Section 3.3.1). In fact, the average state conditional correlations of BTC log-returns and each of the financial features are close to zero in all the three regimes.

3.4.3 Discussion

The SJM model distinguishes three distinct regimes driven by upward, downward and sideways trends, suggesting that time-series momentum is a key driver of cryptocurrencies. Notably, the presence of time-series momentum is well-established in traditional asset classes such as equity, currency, commodity, and fixed income markets (Moskowitz et al., 2012; Babu et al., 2020). More recently it has also been shown to be prevalent in the crypto markets (Cheah et al., 2020; Liu and Tsyvinski, 2021; Liu et al., 2022; Koki et al., 2022).

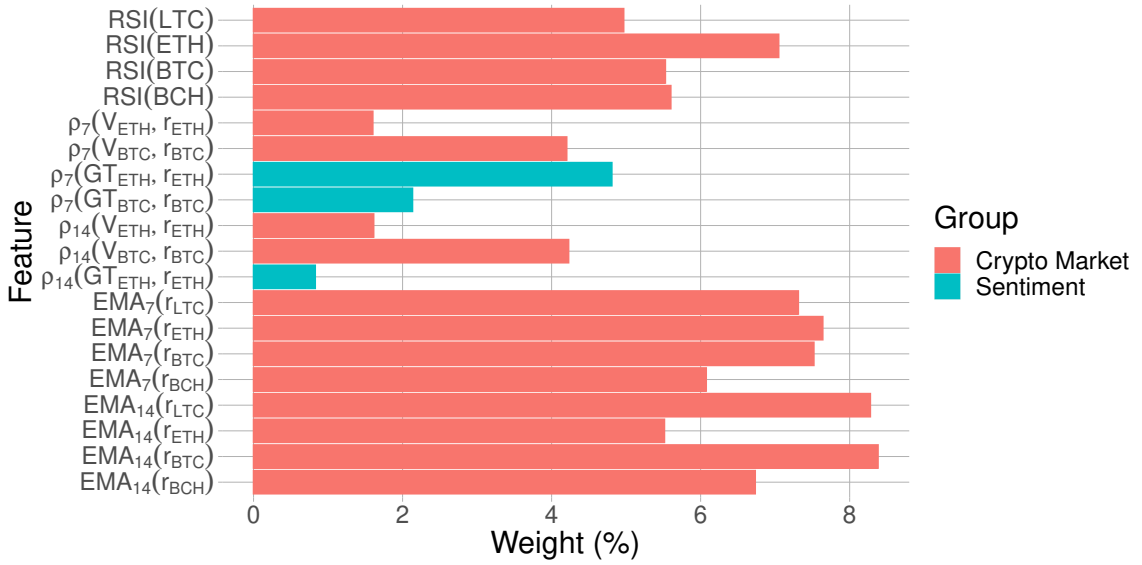


Figure 3.3: Estimated weights of the relevant features in the three-state model. The selected features are RSIs for BTC, ETH, LTC and BCH; 7- and 14-day exponentially weighted linear correlations of log-differences of volumes and log-returns for BTC and ETH; 7-day exponentially weighted linear correlations of log-differences of GT and log-returns for BTC and ETH; 14-day exponentially weighted linear correlations of log-differences of GT and log-returns for BTC; 7- and 14-day EMAs of log-returns of BTC, ETH, LTC and BCH. RSI, ρ_d and EMA_d denote the relative strength index, exponentially weighted linear correlation and exponential moving average with a half-life of d days, respectively.

Theories of sentiment in the behavioral finance literature suggests time-series momentum may result from initial under-reaction followed by delayed over-reaction.[8] Below we provide some support for this explanation and discuss how the interplay between institutional investors, who predominately act as market makers in these markets, and retail investors, who are holding positions longer, contribute to the trends.

While the interest of larger financial institutions in the crypto markets is growing, Karniol-Tambour et al. (2022) estimate that as of January 2022 only around 5% of Bitcoin is *held* by institutional investors. However, although institutions do not appear to have the largest share of holdings in the crypto markets, in the last few years they are generally believed to be the dominant players when it comes to trading volumes (e.g. market making). In 2021, institutions traded over one trillion dollars worth of cryptocurrencies on Coinbase,

[8] Under-reaction is the failure of markets to fully react to new information and can occur through several behavioral channels, including gradual dissemination of news (Hong and Stein, 1999), adherence to prior beliefs and cognitive biases such as anchoring (Barberis et al., 1998), or selling profitable assets too early while holding on to losing ones too long (Shefrin and Statman, 1985). Conversely, over-reaction is the tendency of markets reacting too strongly to new information and can result from excessive optimism and self-attribution biases (Daniel et al., 1998), herding behavior (Bikhchandani et al., 1992), positive feedback trading (De Long et al., 1990), or market sentiment (Baker and Wurgler, 2006).

Table 3.5: Estimated weights and state-conditional values of the selected features. RSI, ρ_d and EMA_d denote the relative strength index, exponentially weighted linear correlation and exponential moving average with a half-life of d days, respectively.

Feature	Weight(%)	Bull	Neutral	Bear
RSI(LTC)	5.0	63.03	43.96	31.81
RSI(ETH)	7.0	66.68	45.03	31.61
RSI(BTC)	5.5	66.46	45.71	31.58
RSI(BCH)	5.6	62.55	43.55	28.60
$\rho_7(V_{\text{ETH}},r_{\text{ETH}})$	1.6	0.21	-0.09	-0.32
$\rho_7(V_{\text{BTC}},r_{\text{BTC}})$	4.2	0.24	-0.07	-0.36
$\rho_7(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	4.8	0.31	-0.08	-0.35
$\rho_7(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	2.1	0.26	-0.10	-0.38
$\rho_{14}(V_{\text{ETH}},r_{\text{ETH}})$	1.6	0.22	-0.06	-0.29
$\rho_{14}(V_{\text{BTC}},r_{\text{BTC}})$	4.2	0.18	-0.04	-0.29
$\rho_{14}(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	0.8	0.14	-0.08	-0.27
$\text{EMA}_7(r_{\text{ETH}})$	7.6	1.08%	-0.25%	1.50%
$\text{EMA}_7(r_{\text{LTC}})$	7.3	0.92%	-0.34%	-1.56%
$\text{EMA}_7(r_{\text{BTC}})$	7.5	0.80%	-0.15%	-1.18%
$\text{EMA}_7(r_{\text{BCH}})$	6.1	0.98%	-0.39%	-2.04%
$\text{EMA}_{14}(r_{\text{ETH}})$	5.5	0.78%	-0.17%	-1.00%
$\text{EMA}_{14}(r_{\text{LTC}})$	8.3	0.65%	-0.27%	-1.12%
$\text{EMA}_{14}(r_{\text{BTC}})$	8.4	0.60%	-0.10%	-0.86%
$\text{EMA}_{14}(r_{\text{BCH}})$	6.7	0.65%	-0.33%	-1.46%

an increase from the 120 billion dollar trading volume the previous year, and more than twice the amount traded by retail investors (half a trillion dollars) (Vigna, 2022). As far as positive momentum, Auer et al. (2022) show that an increase in the price of Bitcoin causes a significant entry of new retail investors in the crypto markets who in turn drive up prices further, consistent with a positive feedback trading explanation (De Long et al., 1990). Using a dataset on client transactions and account balances of retail customers at a large German online bank, Hackethal et al. (2022) suggest that customers investing in cryptocurrencies and cryptocurrency structured retail products are likely to exhibit investing biases consistent with naive trend-chasing and overtrading behavior (Barber and Odean, 2008). Additionally, Kogan et al. (2022) show that many retail investors actually end up following momentum strategies, whether they are aware of it or not, when investing in cryptocurrencies. The authors argue that this behavior is predominantly driven by retail investors holding on to their positions, even in periods of large price moves. In particular, they do not rebalance after prices increase or double up when prices decrease.

The features representing the interaction of cryptocurrency returns with public attention

and trade volumes are also selected by our model. Their state-conditional values are positive in the bull regime, negative in the bear regime, and close to zero in the neutral regime, consistent with, for example, Bianchi and Dickerson (2019); Smales (2022).

Inspecting which groups of features are not selected by the SJM provides additional insight into the workings of crypto markets. Most importantly, features representing traditional asset classes are not helpful in explaining cryptocurrencies (see also Baur et al. (2018); Bianchi (2020)) and neither is cryptocurrency return volatility-based features. That the latter are not selected is perhaps a bit surprising, especially as volatility-based features are some of the most important features for identifying regimes in the equity markets (Nystrup et al., 2020a, 2021). A possible reason for their non-inclusion is that their state-conditional values are about the same in the bull and neutral regimes, with each being about half of the corresponding values in the bear regime.

3.5 Conclusions

We employ the sparse statistical jump model to infer key features that drive the return dynamics of the largest cryptocurrencies. Our results suggest that a model with three states provides an intuitive interpretation of these markets corresponding to bull, neutral and bear market regimes. We find that first moments of returns (but not second moments), features representing trends and reversal signals drawn from the technical analysis literature, market activity and public attention have the strongest descriptive power. The features that we use for representing market activity and public attention are new and aid in explaining cryptocurrency returns in upward and downward market trends.

These findings have practical implications for trading and risk management in the crypto market. In particular, practitioners can use the identified features to distinguish upward and downward market trends, and detect when the market switches between different regimes.

Appendices

3.A Feature set

Tag	Variable(s)	Transformation	Group
r_{BTC}	BTC log-ret	log-difference	Crypto Market-Related
r_{ETH}	ETH log-ret	log-difference	Crypto Market-Related
r_{XRP}	XRP log-ret	log-difference	Crypto Market-Related
r_{LTC}	LTC log-ret	log-difference	Crypto Market-Related
r_{BCH}	BCH log-ret	log-difference	Crypto Market-Related
V_{BTC}	BTC Volume	log-difference	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
V_{ETH}	ETH Volume	log difference	Crypto Market-Related
V_{XRP}	XRP Volume	log-difference	Crypto Market-Related
V_{LTC}	LTC Volume	log-difference	Crypto Market-Related
V_{BCH}	BCH Volume	log-difference	Crypto Market-Related
$\text{EMA}_1(r_{\text{BTC}})$	BTC log-ret	1-day EMA	Crypto Market-Related
$\text{EMA}_1(r_{\text{ETH}})$	ETH log-ret	1-day EMA	Crypto Market-Related
$\text{EMA}_1(r_{\text{XRP}})$	XRP log-ret	1-day EMA	Crypto Market-Related
$\text{EMA}_1(r_{\text{LTC}})$	LTC log-ret	1-day EMA	Crypto Market-Related
$\text{EMA}_1(r_{\text{BCH}})$	BCH log-ret	1-day EMA	Crypto Market-Related
$\text{EMA}_1(\sigma_{\text{BTC}})$	BTC log-ret	1-day EMA volatility	Crypto Market-Related
$\text{EMA}_1(\sigma_{\text{ETH}})$	ETH log-ret	1-day EMA volatility	Crypto Market-Related
$\text{EMA}_1(\sigma_{\text{XRP}})$	XRP log-ret	1-day EMA volatility	Crypto Market-Related
$\text{EMA}_1(\sigma_{\text{LTC}})$	LTC log-ret	1-day EMA volatility	Crypto Market-Related
$\text{EMA}_1(\sigma_{\text{BCH}})$	BCH log-ret	1-day EMA volatility	Crypto Market-Related
$\text{EMA}_2(r_{\text{BTC}})$	BTC log-ret	2-day EMA	Crypto Market-Related
$\text{EMA}_2(r_{\text{ETH}})$	ETH log-ret	2-day EMA	Crypto Market-Related
$\text{EMA}_2(r_{\text{XRP}})$	XRP log-ret	2-day EMA	Crypto Market-Related
$\text{EMA}_2(r_{\text{LTC}})$	LTC log-ret	2-day EMA	Crypto Market-Related
$\text{EMA}_2(r_{\text{BCH}})$	BCH log-ret	2-day EMA	Crypto Market-Related
$\text{EMA}_2(\sigma_{\text{BTC}})$	BTC log-ret	2-day EMA volatility	Crypto Market-Related
$\text{EMA}_2(\sigma_{\text{ETH}})$	ETH log-ret	2-day EMA volatility	Crypto Market-Related
$\text{EMA}_2(\sigma_{\text{XRP}})$	XRP log-ret	2-day EMA volatility	Crypto Market-Related
$\text{EMA}_2(\sigma_{\text{LTC}})$	LTC log-ret	2-day EMA volatility	Crypto Market-Related
$\text{EMA}_2(\sigma_{\text{BCH}})$	BCH log-ret	2-day EMA volatility	Crypto Market-Related
$\text{EMA}_7(r_{\text{BTC}})$	BTC log-ret	7-day EMA	Crypto Market-Related
$\text{EMA}_7(r_{\text{ETH}})$	ETH log-ret	7-day EMA	Crypto Market-Related
$\text{EMA}_7(r_{\text{XRP}})$	XRP log-ret	7-day EMA	Crypto Market-Related
$\text{EMA}_7(r_{\text{LTC}})$	LTC log-ret	7-day EMA	Crypto Market-Related
$\text{EMA}_7(r_{\text{BCH}})$	BCH log-ret	7-day EMA	Crypto Market-Related
$\text{EMA}_7(\sigma_{\text{BTC}})$	BTC log-ret	7-day EMA volatility	Crypto Market-Related
$\text{EMA}_7(\sigma_{\text{ETH}})$	ETH log-ret	7-day EMA volatility	Crypto Market-Related
$\text{EMA}_7(\sigma_{\text{XRP}})$	XRP log-ret	7-day EMA volatility	Crypto Market-Related
$\text{EMA}_7(\sigma_{\text{LTC}})$	LTC log-ret	7-day EMA volatility	Crypto Market-Related
$\text{EMA}_7(\sigma_{\text{BCH}})$	BCH log-ret	7-day EMA volatility	Crypto Market-Related
$\text{EMA}_{14}(r_{\text{BTC}})$	BTC log-ret	14-day EMA	Crypto Market-Related
$\text{EMA}_{14}(r_{\text{ETH}})$	ETH log-ret	14-day EMA	Crypto Market-Related
$\text{EMA}_{14}(r_{\text{XRP}})$	XRP log-ret	14-day EMA	Crypto Market-Related
$\text{EMA}_{14}(r_{\text{LTC}})$	LTC log-ret	14-day EMA	Crypto Market-Related
$\text{EMA}_{14}(r_{\text{BCH}})$	BCH log-ret	14-day EMA	Crypto Market-Related
$\text{EMA}_{14}(\sigma_{\text{BTC}})$	BTC log-ret	14-day EMA volatility	Crypto Market-Related
$\text{EMA}_{14}(\sigma_{\text{ETH}})$	ETH log-ret	14-day EMA volatility	Crypto Market-Related
$\text{EMA}_{14}(\sigma_{\text{XRP}})$	XRP log-ret	14-day EMA volatility	Crypto Market-Related
$\text{EMA}_{14}(\sigma_{\text{LTC}})$	LTC log-ret	14-day EMA volatility	Crypto Market-Related
$\text{EMA}_{14}(\sigma_{\text{BCH}})$	BCH log-ret	14-day EMA volatility	Crypto Market-Related
$\text{EMA}_1(V_{\text{BTC}})$	BTC Volume	1-day EMA Volume	Crypto Market-Related
$\text{EMA}_1(V_{\text{ETH}})$	ETH Volume	1-day EMA Volume	Crypto Market-Related
$\text{EMA}_1(V_{\text{XRP}})$	XRP Volume	1-day EMA Volume	Crypto Market-Related
$\text{EMA}_1(V_{\text{LTC}})$	LTC Volume	1-day EMA Volume	Crypto Market-Related
$\text{EMA}_1(V_{\text{BCH}})$	BCH Volume	1-day EMA Volume	Crypto Market-Related
$\text{EMA}_2(V_{\text{BTC}})$	BTC Volume	2-day EMA Volume	Crypto Market-Related
$\text{EMA}_2(V_{\text{ETH}})$	ETH Volume	2-day EMA Volume	Crypto Market-Related
$\text{EMA}_2(V_{\text{XRP}})$	XRP Volume	2-day EMA Volume	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
$\text{EMA}_2(V_{\text{LTC}})$	LTC Volume	2-day EMA Volume	Crypto Market-Related
$\text{EMA}_2(V_{\text{BCH}})$	BCH Volume	2-day EMA Volume	Crypto Market-Related
$\text{EMA}_7(V_{\text{BTC}})$	BTC Volume	7-day EMA Volume	Crypto Market-Related
$\text{EMA}_7(V_{\text{ETH}})$	ETH Volume	7-day EMA Volume	Crypto Market-Related
$\text{EMA}_7(V_{\text{XRP}})$	XRP Volume	7-day EMA Volume	Crypto Market-Related
$\text{EMA}_7(V_{\text{LTC}})$	LTC Volume	7-day EMA Volume	Crypto Market-Related
$\text{EMA}_7(V_{\text{BCH}})$	BCH Volume	7-day EMA Volume	Crypto Market-Related
$\text{EMA}_{14}(V_{\text{BTC}})$	BTC Volume	14-day EMA Volume	Crypto Market-Related
$\text{EMA}_{14}(V_{\text{ETH}})$	ETH Volume	14-day EMA Volume	Crypto Market-Related
$\text{EMA}_{14}(V_{\text{XRP}})$	XRP Volume	14-day EMA Volume	Crypto Market-Related
$\text{EMA}_{14}(V_{\text{LTC}})$	LTC Volume	14-day EMA Volume	Crypto Market-Related
$\text{EMA}_{14}(V_{\text{BCH}})$	BCH Volume	14-day EMA Volume	Crypto Market-Related
$\rho_1(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	1-day EMA linear correlation	Crypto Market-Related
$\rho_2(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	2-day EMA linear correlation	Crypto Market-Related
$\rho_7(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	7-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	14-day EMA linear correlation	Crypto Market-Related
$g_1(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	1-day EMA Gerber correlation	Crypto Market-Related
$g_2(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	2-day EMA Gerber correlation	Crypto Market-Related
$g_7(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	7-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(V_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Volume	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(V_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Volume	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(V_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Volume	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(V_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Volume	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(V_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Volume	14-day EMA Gerber correlation	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
VOC_{BTC}	BTC on chain volume	first difference	Crypto Market-Related
VOC_{ETH}	ETH on chain volume	first difference	Crypto Market-Related
VOC_{LTC}	LTC on chain volume	first difference	Crypto Market-Related
VOC_{BCH}	BCH on chain volume	first difference	Crypto Market-Related
HR_{BTC}	BTC hash-rate	log-difference	Crypto Market-Related
HR_{ETH}	ETH hash-rate	log-difference	Crypto Market-Related
$\text{AddWB}_{\text{BTC}}$	BTC number of total addresses with balance	log-difference	Crypto Market-Related
$\text{AddWB}_{\text{ETH}}$	ETH number of total addresses with balance	log-difference	Crypto Market-Related
$\text{AddWB}_{\text{LTC}}$	LTC number of total addresses with balance	log-difference	Crypto Market-Related
$\text{AddWB}_{\text{BCH}}$	BCH number of total addresses with balance	log-difference	Crypto Market-Related
$\rho_1(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	1-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	2-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	7-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	14-day EMA linear correlation	Crypto Market-Related
$g_1(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	1-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	2-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	7-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{VOC}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC volume on chain	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{VOC}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET volume on chain	14-day EMA Gerber correlation	Crypto Market-Related
$\rho_1(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	1-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	2-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	7-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	14-day EMA linear correlation	Crypto Market-Related
$g_1(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	1-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	2-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	7-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{VOC}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC volume on chain	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{VOC}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret ET volume on chain	14-day EMA Gerber correlation	Crypto Market-Related
$\rho_1(\text{AddWB}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(\text{AddWB}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(\text{AddWB}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	1-day EMA linear correlation	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
$\rho_1(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	1-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	2-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	7-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	14-day EMA linear correlation	Crypto Market-Related
$g_1(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	1-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	2-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{AddWB}_{\text{ETH}}, r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{AddWB}_{\text{LTC}}, r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	7-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{AddWB}_{\text{BTC}}, r_{\text{BTC}})$	BTC log-ret BTC number of total addresses with balance with balance	14-day EMA Gerber correlation	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
$g_{14}(\text{AddWB}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET number of total addresses with balance with balance	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{AddWB}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC number of total addresses with balance with balance	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{AddWB}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH number of total addresses with balance with balance	14-day EMA Gerber correlation	Crypto Market-Related
$\rho_1(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	1-day EMA linear correlation	Crypto Market-Related
$\rho_1(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	1-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	2-day EMA linear correlation	Crypto Market-Related
$\rho_2(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	2-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	7-day EMA linear correlation	Crypto Market-Related
$\rho_7(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	7-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	14-day EMA linear correlation	Crypto Market-Related
$\rho_{14}(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	14-day EMA linear correlation	Crypto Market-Related
$g_1(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	1-day EMA Gerber correlation	Crypto Market-Related
$g_1(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	1-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	2-day EMA Gerber correlation	Crypto Market-Related
$g_2(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	2-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	7-day EMA Gerber correlation	Crypto Market-Related
$g_7(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	7-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{HR}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC hash rate	14-day EMA Gerber correlation	Crypto Market-Related
$g_{14}(\text{HR}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET hash rate	14-day EMA Gerber correlation	Crypto Market-Related
RF_1(BTC)	BTC log-ret BTC Volatility	time-series regression forecast l=1	Crypto Market-Related
RF_2(BTC)	BTC log-ret BTC Volatility	time-series regression forecast l=2	Crypto Market-Related
RF_7(BTC)	BTC log-ret BTC Volatility	time-series regression forecast l=7	Crypto Market-Related
RF_{14}(BTC)	BTC log-ret BTC Volatility	time-series regression forecast l=14	Crypto Market-Related
RF_1(ETH)	ETH log-ret ETH Volatility	time-series regression forecast l=1	Crypto Market-Related
RF_2(ETH)	ETH log-ret ETH Volatility	time-series regression forecast l=2	Crypto Market-Related
RF_7(ETH)	ETH log-ret ETH Volatility	time-series regression forecast l=7	Crypto Market-Related
RF_{14}(ETH)	ETH log-ret ETH Volatility	time-series regression forecast l=14	Crypto Market-Related
RF_1(XRP)	XRP log-ret XRP Volatility	time-series regression forecast l=1	Crypto Market-Related
RF_2(XRP)	XRP log-ret XRP Volatility	time-series regression forecast l=2	Crypto Market-Related
RF_7(XRP)	XRP log-ret XRP Volatility	time-series regression forecast l=7	Crypto Market-Related
RF_{14}(XRP)	XRP log-ret XRP Volatility	time-series regression forecast l=14	Crypto Market-Related
RF_1(LTC)	LTC log-ret LTC Volatility	time-series regression forecast l=1	Crypto Market-Related
RF_2(LTC)	LTC log-ret LTC Volatility	time-series regression forecast l=2	Crypto Market-Related
RF_7(LTC)	LTC log-ret LTC Volatility	time-series regression forecast l=7	Crypto Market-Related
RF_{14}(LTC)	LTC log-ret LTC Volatility	time-series regression forecast l=14	Crypto Market-Related
RF_1(BCH)	BCH log-ret BCH Volatility	time-series regression forecast l=1	Crypto Market-Related
RF_2(BCH)	BCH log-ret BCH Volatility	time-series regression forecast l=2	Crypto Market-Related
RF_7(BCH)	BCH log-ret BCH Volatility	time-series regression forecast l=7	Crypto Market-Related
RF_{14}(BCH)	BCH log-ret BCH Volatility	time-series regression forecast l=14	Crypto Market-Related
RSI(BTC)	BTC Price	RSI	Crypto Market-Related
MACDS(BTC)	BTC Price	MACD minus signal	Crypto Market-Related
RSI(ETH)	ETH Price	RSI	Crypto Market-Related
MACDS(ETH)	ETH Price	MACD minus signal	Crypto Market-Related
RSI(XRP)	XRP Price	RSI	Crypto Market-Related
MACDS(XRP)	XRP Price	MACD minus signal	Crypto Market-Related
RSI(LTC)	LTC Price	RSI	Crypto Market-Related
MACDS(LTC)	LTC Price	MACD minus signal	Crypto Market-Related
RSI(BCH)	BCH Price	RSI	Crypto Market-Related
MACDS(BCH)	BCH Price	MACD minus signal	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
$\rho_1(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	1-day linear correlation	Crypto Market-Related
$\rho_1(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	1-day linear correlation	Crypto Market-Related
$\rho_1(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	1-day linear correlation	Crypto Market-Related
$\rho_1(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	1-day linear correlation	Crypto Market-Related
$\rho_2(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	2-day linear correlation	Crypto Market-Related
$\rho_2(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	2-day linear correlation	Crypto Market-Related
$\rho_2(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	2-day linear correlation	Crypto Market-Related
$\rho_2(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	2-day linear correlation	Crypto Market-Related
$\rho_7(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	7-day linear correlation	Crypto Market-Related
$\rho_7(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	7-day linear correlation	Crypto Market-Related
$\rho_7(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	7-day linear correlation	Crypto Market-Related
$\rho_7(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	7-day linear correlation	Crypto Market-Related
$\rho_{14}(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	14-day linear correlation	Crypto Market-Related
$\rho_{14}(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	14-day linear correlation	Crypto Market-Related
$\rho_{14}(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	14-day linear correlation	Crypto Market-Related
$\rho_{14}(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	14-day linear correlation	Crypto Market-Related
$g_1(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	1-day linear correlation	Crypto Market-Related
$g_1(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	1-day linear correlation	Crypto Market-Related
$g_1(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	1-day linear correlation	Crypto Market-Related
$g_1(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	1-day linear correlation	Crypto Market-Related
$g_2(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	2-day linear correlation	Crypto Market-Related
$g_2(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	2-day linear correlation	Crypto Market-Related
$g_2(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	2-day linear correlation	Crypto Market-Related
$g_2(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	2-day linear correlation	Crypto Market-Related
$g_7(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	7-day linear correlation	Crypto Market-Related
$g_7(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	7-day linear correlation	Crypto Market-Related
$g_7(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	7-day linear correlation	Crypto Market-Related
$g_7(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	7-day linear correlation	Crypto Market-Related
$g_{14}(r_{\text{ETH}},r_{\text{BTC}})$	BTC log-ret ETH log-ret	14-day linear correlation	Crypto Market-Related
$g_{14}(r_{\text{XRP}},r_{\text{BTC}})$	BTC log-ret XRP log-ret	14-day linear correlation	Crypto Market-Related
$g_{14}(r_{\text{LTC}},r_{\text{BTC}})$	BTC log-ret LTC log-ret	14-day linear correlation	Crypto Market-Related
$g_{14}(r_{\text{BCH}},r_{\text{BTC}})$	BTC log-ret BCH log-ret	14-day linear correlation	Crypto Market-Related
TM(BTC)	BTC on chain volume with balance BTC prices	first difference of ratio	Crypto Market-Related
TM(ETH)	ETH on chain volume with balance ETH prices	first difference of ratio	Crypto Market-Related
TM(LTC)	LTC on chain volume with balance LTC prices	first difference of ratio	Crypto Market-Related
TM(BCH)	BCH number of total addresses with balance BCH prices	first difference of ratio	Crypto Market-Related
AM(BTC)	BTC number of total addresses BTC prices	log-difference of ratio	Crypto Market-Related
AM(ETH)	ETH number of total addresses ETH prices	log-difference of ratio	Crypto Market-Related
AM(LTC)	LTC number of total addresses LTC prices	log-difference of ratio	Crypto Market-Related
AM(BCH)	BCH nuber of total addresses BCH prices	log-difference of ratio	Crypto Market-Related
UM(BTC)	BTC number of total addresses with balance BTC prices	log-difference of ratio	Crypto Market-Related
UM(ETH)	ETH number of total addresses with balance ETH prices	log-difference of ratio	Crypto Market-Related

Continued on next page

Tag	Variable(s)	Transformation	Group
UM(LTC)	LTC number of total addresses with balance LTC prices	log-difference of ratio	Crypto Market-Related
UM(BCH)	BCH nuber of total addresses with balance BCH prices	log-difference of ratio	Crypto Market-Related
$\text{AMIHUD}_{\text{BTC}}$	BTC absolute log-ret BTC volume	ratio	Crypto Market-Related
$\text{AMIHUD}_{\text{ETH}}$	ETH absolute log-ret ETH volume	ratio	Crypto Market-Related
$\text{AMIHUD}_{\text{XRP}}$	XRP absolute log-ret XRP volume	ratio	Crypto Market-Related
$\text{AMIHUD}_{\text{LTC}}$	LTC absolute log-ret LTC volume	ratio	Crypto Market-Related
$\text{AMIHUD}_{\text{BCH}}$	BCH absolute log-ret BCH volume	ratio	Crypto Market-Related
GT_{BTC}	BTC Google Index	first difference	Sentiment
GT_{ETH}	ETH Google Index	first difference	Sentiment
GT_{XRP}	XRP Google Index	first difference	Sentiment
GT_{LTC}	LTC Google Index	first difference	Sentiment
GT_{BCH}	BCH Google Index	first difference	Sentiment
$\rho_1(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	1-day EMA linear correlation	Sentiment
$\rho_1(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	1-day EMA linear correlation	Sentiment
$\rho_1(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	1-day EMA linear correlation	Sentiment
$\rho_1(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	1-day EMA linear correlation	Sentiment
$\rho_1(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	1-day EMA linear correlation	Sentiment
$\rho_2(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	2-day EMA linear correlation	Sentiment
$\rho_2(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	2-day EMA linear correlation	Sentiment
$\rho_2(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	2-day EMA linear correlation	Sentiment
$\rho_2(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	2-day EMA linear correlation	Sentiment
$\rho_2(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	2-day EMA linear correlation	Sentiment
$\rho_7(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	7-day EMA linear correlation	Sentiment
$\rho_7(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	7-day EMA linear correlation	Sentiment
$\rho_7(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	7-day EMA linear correlation	Sentiment
$\rho_7(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	7-day EMA linear correlation	Sentiment
$\rho_7(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	7-day EMA linear correlation	Sentiment
$\rho_{14}(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	14-day EMA linear correlation	Sentiment
$\rho_{14}(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	14-day EMA linear correlation	Sentiment
$\rho_{14}(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	14-day EMA linear correlation	Sentiment
$\rho_{14}(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	14-day EMA linear correlation	Sentiment
$\rho_{14}(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	14-day EMA linear correlation	Sentiment
$g_1(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	1-day EMA Gerber correlation	Sentiment
$g_1(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	1-day EMA Gerber correlation	Sentiment
$g_1(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	1-day EMA Gerber correlation	Sentiment
$g_1(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	1-day EMA Gerber correlation	Sentiment
$g_1(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	1-day EMA Gerber correlation	Sentiment
$g_2(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	2-day EMA Gerber correlation	Sentiment
$g_2(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	2-day EMA Gerber correlation	Sentiment
$g_2(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	2-day EMA Gerber correlation	Sentiment
$g_2(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	2-day EMA Gerber correlation	Sentiment
$g_2(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	2-day EMA Gerber correlation	Sentiment
$g_7(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	7-day EMA Gerber correlation	Sentiment
$g_7(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	7-day EMA Gerber correlation	Sentiment
$g_7(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	7-day EMA Gerber correlation	Sentiment
$g_7(\text{GT}_{\text{LTC}},r_{\text{LTC}})$	LTC log-ret LTC Google	7-day EMA Gerber correlation	Sentiment
$g_7(\text{GT}_{\text{BCH}},r_{\text{BCH}})$	BCH log-ret BCH Google	7-day EMA Gerber correlation	Sentiment
$g_{14}(\text{GT}_{\text{BTC}},r_{\text{BTC}})$	BTC log-ret BTC Google	14-day EMA Gerber correlation	Sentiment
$g_{14}(\text{GT}_{\text{ETH}},r_{\text{ETH}})$	ETH log-ret ET Google	14-day EMA Gerber correlation	Sentiment
$g_{14}(\text{GT}_{\text{XRP}},r_{\text{XRP}})$	XRP log-ret XRP Google	14-day EMA Gerber correlation	Sentiment

Continued on next page

Tag	Variable(s)	Transformation	Group
$g_{14}(\mathrm{GT}_{\mathrm{LTC}},r_{\mathrm{LTC}})$	LTC log-ret LTC Google	14-day EMA Gerber correlation	Sentiment
$g_{14}(\mathrm{GT}_{\mathrm{BCH}},r_{\mathrm{BCH}})$	BCH log-ret BCH Google	14-day EMA Gerber correlation	Sentiment
GOLD	Gold log-ret	log-difference	Financial Market
EURUSD	EURUSD log-ret	log-difference	Financial Market
JPYUSD	JPYUSD log-ret	log-difference	Financial Market
CNYUSD	CNYUSD log-ret	log-difference	Financial Market
VIX	VIX log-ret	log-difference	Financial Market
SP500	SP500 log-ret	log-difference	Financial Market
NASDAQ	NASDAQ log-ret	log-difference	Financial Market
T10Y3M	10 years minus 3 months US treasury yields	first difference	Financial Market
WTI	WTI	first difference	Financial Market
$\rho_1(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	1-day EMA Gerber correlation	Financial Market
$\rho_2(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	2-day EMA Gerber correlation	Financial Market
$\rho_7(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	7-day EMA Gerber correlation	Financial Market
$\rho_{14}(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	14-day EMA Gerber correlation	Financial Market
$g_1(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	1-day EMA Gerber correlation	Financial Market
$g_2(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	2-day EMA Gerber correlation	Financial Market
$g_7(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	7-day EMA Gerber correlation	Financial Market
$g_{14}(\mathrm{GOLD},r_{\mathrm{BTC}})$	BTC log-ret gold log-ret	14-day EMA Gerber correlation	Financial Market
$\rho_1(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	1-day EMA Gerber correlation	Financial Market
$\rho_2(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	2-day EMA Gerber correlation	Financial Market
$\rho_7(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	7-day EMA Gerber correlation	Financial Market
$\rho_{14}(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	14-day EMA Gerber correlation	Financial Market
$g_1(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	1-day EMA Gerber correlation	Financial Market
$g_2(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	2-day EMA Gerber correlation	Financial Market
$g_7(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	7-day EMA Gerber correlation	Financial Market
$g_{14}(\mathrm{EURUSD},r_{\mathrm{BTC}})$	BTC log-ret EURUSD log-ret	14-day EMA Gerber correlation	Financial Market
$\rho_1(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	1-day EMA Gerber correlation	Financial Market
$\rho_2(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	2-day EMA Gerber correlation	Financial Market
$\rho_7(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	7-day EMA Gerber correlation	Financial Market
$\rho_{14}(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	14-day EMA Gerber correlation	Financial Market
$g_1(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	1-day EMA Gerber correlation	Financial Market
$g_2(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	2-day EMA Gerber correlation	Financial Market
$g_7(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	7-day EMA Gerber correlation	Financial Market
$g_{14}(\mathrm{JPYUSD},r_{\mathrm{BTC}})$	BTC log-ret JPYUSD log-ret	14-day EMA Gerber correlation	Financial Market
$\rho_1(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	1-day EMA Gerber correlation	Financial Market
$\rho_2(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	2-day EMA Gerber correlation	Financial Market
$\rho_7(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	7-day EMA Gerber correlation	Financial Market
$\rho_{14}(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	14-day EMA Gerber correlation	Financial Market
$g_1(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	1-day EMA Gerber correlation	Financial Market
$g_2(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	2-day EMA Gerber correlation	Financial Market
$g_7(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	7-day EMA Gerber correlation	Financial Market
$g_{14}(\mathrm{CNYUSD},r_{\mathrm{BTC}})$	BTC log-ret CNYUSD log-ret	14-day EMA Gerber correlation	Financial Market
$\rho_1(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	1-day EMA Gerber correlation	Financial Market
$\rho_2(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	2-day EMA Gerber correlation	Financial Market
$\rho_7(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	7-day EMA Gerber correlation	Financial Market
$\rho_{14}(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	14-day EMA Gerber correlation	Financial Market
$g_1(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	1-day EMA Gerber correlation	Financial Market
$g_2(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	2-day EMA Gerber correlation	Financial Market
$g_7(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	7-day EMA Gerber correlation	Financial Market
$g_{14}(\mathrm{VIX},r_{\mathrm{BTC}})$	BTC log-ret VIX log-ret	14-day EMA Gerber correlation	Financial Market

Continued on next page

Tag	Variable(s)	Transformation	Group
ρ_1(SP500,r_{BTC})	BTC log-ret SP500 log-ret	1-day EMA Gerber correlation	Financial Market
ρ_2(SP500,r_{BTC})	BTC log-ret SP500 log-ret	2-day EMA Gerber correlation	Financial Market
ρ_7(SP500,r_{BTC})	BTC log-ret SP500 log-ret	7-day EMA Gerber correlation	Financial Market
ρ_{14}(SP500,r_{BTC})	BTC log-ret SP500 log-ret	14-day EMA Gerber correlation	Financial Market
g_1(SP500,r_{BTC})	BTC log-ret SP500 log-ret	1-day EMA Gerber correlation	Financial Market
g_2(SP500,r_{BTC})	BTC log-ret SP500 log-ret	2-day EMA Gerber correlation	Financial Market
g_7(SP500,r_{BTC})	BTC log-ret SP500 log-ret	7-day EMA Gerber correlation	Financial Market
g_{14}(SP500,r_{BTC})	BTC log-ret SP500 log-ret	14-day EMA Gerber correlation	Financial Market
ρ_1(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	1-day EMA Gerber correlation	Financial Market
ρ_2(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	2-day EMA Gerber correlation	Financial Market
ρ_7(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	7-day EMA Gerber correlation	Financial Market
ρ_{14}(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	14-day EMA Gerber correlation	Financial Market
g_1(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	1-day EMA Gerber correlation	Financial Market
g_2(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	2-day EMA Gerber correlation	Financial Market
g_7(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	7-day EMA Gerber correlation	Financial Market
g_{14}(NASDAQ,r_{BTC})	BTC log-ret NASDAQ log-ret	14-day EMA Gerber correlation	Financial Market
ρ_1(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	1-day EMA Gerber correlation	Financial Market
ρ_2(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	2-day EMA Gerber correlation	Financial Market
ρ_7(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	7-day EMA Gerber correlation	Financial Market
ρ_{14}(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	14-day EMA Gerber correlation	Financial Market
g_1(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	1-day EMA Gerber correlation	Financial Market
g_2(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	2-day EMA Gerber correlation	Financial Market
g_7(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	7-day EMA Gerber correlation	Financial Market
g_{14}(T10Y3M, r_{BTC})	BTC log-ret T10Y3M	14-day EMA Gerber correlation	Financial Market
ρ_1(WTI,r_{BTC})	BTC log-ret WTI	1-day EMA Gerber correlation	Financial Market
ρ_2(WTI,r_{BTC})	BTC log-ret WTI	2-day EMA Gerber correlation	Financial Market
ρ_7(WTI,r_{BTC})	BTC log-ret WTI	7-day EMA Gerber correlation	Financial Market
ρ_{14}(WTI,r_{BTC})	BTC log-ret WTI	14-day EMA Gerber correlation	Financial Market
g_1(WTI,r_{BTC})	BTC log-ret WTI	1-day EMA Gerber correlation	Financial Market
g_2(WTI,r_{BTC})	BTC log-ret WTI	2-day EMA Gerber correlation	Financial Market
g_7(WTI,r_{BTC})	BTC log-ret WTI	7-day EMA Gerber correlation	Financial Market
g_{14}(WTI,r_{BTC})	BTC log-ret WTI	14-day EMA Gerber correlation	Financial Market
EMA_1(VIX)	VIX log-ret	1-day EMA	Financial Market
EMA_2(VIX)	VIX log-ret	2-day EMA	Financial Market
EMA_7(VIX)	VIX log-ret	7-day EMA	Financial Market
EMA_{14}(VIX)	VIX log-ret	14-day EMA	Financial Market

Bibliography

Aalborg, H. A., Molnár, P., and de Vries, J. E. (2019). What can explain the price, volatility and trading volume of Bitcoin? *Finance Research Letters*, 29:255–265.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Alexander, C. and Dakos, M. (2020). A critical investigation of cryptocurrency data and analysis. *Quantitative Finance*, 20:173–188.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5:31–56.

Appel, G. (2005). *Technical Analysis: Power Tools for Active Investors*. FT Press.

Ardia, D., Bluteau, K., and Rüede, M. (2019). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29:266–271.

Auer, R., Cornelli, G., Doerr, S., Frost, J., Gambacorta, L., et al. (2022). Crypto trading and Bitcoin prices: Evidence from a new database of retail adoption. Technical report, Bank for International Settlements.

Babu, A., Levine, A., Ooi, Y. H., Pedersen, L. H., and Stamelos, E. (2020). Trends everywhere. *Journal of Investment Management*, 18:52–68.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61:1645–1680.

Baker, S. and Bloom, N. (2013). Does uncertainty reduce growth? Using disasters as natural experiments. Technical report, National Bureau of Economic Research.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21:785–818.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49:307–343.

Barucci, E., Moncayo, G. G., and Marazzina, D. (2022). Cryptocurrencies and stablecoins: A high-frequency analysis. *Digital Finance*, 4:217–239.

Baur, D. G., Hong, K., and Lee, A. D. (2018). Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, 54:177–189.

Bemporad, A., Breschi, V., Piga, D., and Boyd, S. P. (2018). Fitting jump models. *Automatica*, 96:11–21.

Bianchi, D. (2020). Cryptocurrencies as an asset class? An empirical assessment. *The Journal of Alternative Investments*, 23:162–179.

Bianchi, D. and Dickerson, A. (2019). Trading volume in cryptocurrency markets. *Available at SSRN 3239670*.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100:992–1026.

Bulla, J. (2011). Hidden Markov models with t components. Increased persistence and other aspects. *Quantitative Finance*, 11:459–475.

Cappiello, L., Engle, R. F., and Sheppard, K. (2006). Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics*, 4:537–572.

Catania, L., Grassi, S., and Ravazzolo, F. (2019). Forecasting cryptocurrencies under model and parameter instability. *International Journal of Forecasting*, 35:485–501.

Chaim, P. and Laurini, M. P. (2018). Volatility and return jumps in Bitcoin. *Economics Letters*, 173:158–163.

Cheah, J. E.-T., Luo, D., Zhang, Z., and Sung, M.-C. (2020). Predictability of Bitcoin returns. *The European Journal of Finance*, 28:66–85.

Cong, L. W., Karolyi, G. A., Tang, K., and Zhao, W. (2021). Value premium, network adoption, and factor pricing of crypto assets. *Working Paper*.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1:223.

Cortese, F., Kolm, P. N., and Lindström, E. (2023). Generalized information criteria for sparse statistical jump models. In Linde, P., editor, *Symposium i Anvendt Statistik*, volume 44. Copenhagen Business School, Copenhagen.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *The Journal of Finance*, 53:1839–1885.

De Bandt, O. and Hartmann, P. (2000). Systemic risk: A survey. *European Central Bank Working Paper*.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance*, 45:379–395.

Erb, C. B., Harvey, C. R., and Viskanta, T. E. (1994). Forecasting international equity correlations. *Financial Analysts Journal*, 50:32–45.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:531–552.

Figà-Talamanca, G., Focardi, S., and Patacca, M. (2021). Regime switches and commonalities of the cryptocurrencies asset class. *The North American Journal of Economics and Finance*, 57:101425.

Figa-Talamanca, G. and Patacca, M. (2019). Does market attention affect Bitcoin returns and volatility? *Decisions in Economics and Finance*, 42:135–155.

Gerber, S., Markowitz, H. M., Ernst, P. A., Miao, Y., Javid, B., and Sargen, P. (2022). The Gerber statistic: A robust co-movement measure for portfolio optimization. *The Journal of Portfolio Management*, 48:87–102.

Hackethal, A., Hanspal, T., Lammer, D. M., and Rink, K. (2022). The characteristics and portfolio behavior of Bitcoin investors: Evidence from indirect cryptocurrency investments. *Review of Finance*, 26:855–898.

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

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance*, 54:2143–2184.

Karniol-Tambour, K., Tan, R., Tsarapkina, D., Sondheimer, J., and Barnes, W. (2022). The evolution of institutional investors' exposure to cryptocurrencies and blockchain technologies. Technical report, Bridgewater Associates, LP.

Kogan, S., Makarov, I., Niessner, M., and Schoar, A. (2022). Are cryptos different? Evidence from retail trading. *Available at SSRN 4289513*.

Koki, C., Leonardos, S., and Piliouras, G. (2022). Exploring the predictability of cryptocurrencies via Bayesian hidden Markov models. *Research in International Business and Finance*, 59:101554.

Kristoufek, L. (2013). Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific Reports*, 3:3415.

Kristoufek, L. (2015). What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PloS One*, 10:e0123923.

Lindström, E., Madsen, H., and Nielsen, J. N. (2015). *Statistics for Finance: Texts in Statistical Science*. Chapman and Hall/CRC.

Liu, Y. and Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *The Review of Financial Studies*, 34:2689–2727.

Liu, Y., Tsyvinski, A., and Wu, X. (2022). Common risk factors in cryptocurrency. *The Journal of Finance*, 77:1133–1177.

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). Time-series momentum. *Journal of Financial Economics*, 104:228–250.

Nystrup, P., Kolm, P. N., and Lindström, E. (2020a). Greedy online classification of persistent market states using realized intraday volatility features. *The Journal of Financial Data Science*, 2:25–39.

Nystrup, P., Kolm, P. N., and Lindström, E. (2021). Feature selection in jump models. *Expert Systems with Applications*, 184:115558.

Nystrup, P., Lindström, E., and Madsen, H. (2020b). Learning hidden Markov models with persistent states by penalizing jumps. *Expert Systems with Applications*, 150:113307.

Pele, D. T., Wesselhöfft, N., Härdle, W. K., Kolossiatis, M., and Yatracos, Y. G. (2021). Are cryptos becoming alternative assets? *The European Journal of Finance*, 29:1064–1105.

Pennoni, F., Bartolucci, F., Forte, G., and Ametrano, F. (2021). Exploring the dependencies among main cryptocurrency log-returns: A hidden Markov model. *Economic Notes*, 51:e12193.

Rydén, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3:659–688.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, 13:217–244.

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

Selmi, R., Mensi, W., Hammoudeh, S., and Bouoiyour, J. (2018). Is Bitcoin a hedge, a safe haven or a diversifier for oil price movements? A comparison with gold. *Energy Economics*, 74:787–801.

Shefrin, H. and Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, 40:777–790.

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

Smales, L. A. (2022). Investor attention in cryptocurrency markets. *International Review of Financial Analysis*, 79:101972.

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

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67.

Vigna, P. (2022). Wall street takes lead in crypto investments. *The Wall Street Journal*.

Wilder, J. W. (1978). *New Concepts in Technical Trading Systems*. Trend Research.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105:713–726.

Xiong, J., Liu, Q., and Zhao, L. (2020). A new method to verify Bitcoin bubbles based on the production cost. *North American Journal of Economics and Finance*, 51:101095.

Yae, J. and Tian, G. Z. (2022). Out-of-sample forecasting of cryptocurrency returns: A comprehensive comparison of predictors and algorithms. *Physica A: Statistical Mechanics and its Applications*, 598:127379.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R*. CRC press, Boca Raton, FL.

Chapter 4

Generalized information criteria for sparse statistical jump models[1]

4.1 Introduction

Regime switching models, such as hidden Markov models (HMMs), are widely employed when data exhibits structural breaks. Common applications include natural language processing, signal processing, wind and solar power forecasting, biology and finance (see Bartolucci et al. (2013) or Zucchini et al. (2017) for an overview). HMMs effectively capture stylized facts of financial markets, especially when parameters are allowed to be time-varying (Nystrup et al., 2017). However, statistical inference poses challenges due to the complexity of optimizing the log-likelihood function, sensitivity to model misspecification, and suboptimal initialization (Rydén et al., 1998; Rydén, 2008; Nystrup et al., 2020c).

Introduced by Bemporad et al. (2018), the class of so-called *statistical jump models* (JMs) offers an intriguing alternative to HMMs. It provides a framework for modeling complex dynamics by transitioning (*"jumping"*) between simpler models. They demonstrate that a standard Gaussian HMM is a special case within their framework. Nystrup et al. (2020a) and Nystrup et al. (2020c) build upon this work by using K-means clustering for the local models. The resulting algorithm conducts temporal clustering, where the cluster persistence is controlled by a hyperparameter. Nystrup et al. (2020a) apply this framework to streaming data, while Nystrup et al. (2021) propose a feature selection framework for this form of temporal clustering, referred to as *sparse statistical jump models* (SJMs). Unlike classical HMMs, the fitting algorithm for SJMs converges quickly, even when there is a substantial number of irrelevant features. It is also robust towards model misspecification and poor initialization. Additionally, SJMs can seamlessly integrate exogenous variables, typically a challenging aspect of classical HMMs (Rydén, 2008; Bartolucci et al., 2014).

[1] A preliminary version of this Chapter has been published in: Cortese F., Kolm P., Lindström E. (2023). Generalized information criteria for sparse statistical jump models. In Linde, P., editor, *Symposium i Anvendt Statistik, volume 44*. www.statistiksymposium.dk

An open question in the practical application of SJMs is the need to properly choose the hyperparameters that govern their behavior. Specifically, these hyperparameters relate to the temporal persistence of the latent states, the selection of relevant features, and the number of clusters used to model the data. Manually tuning these hyperparameters can be a challenging and time-consuming task, and it can be difficult to obtain optimal results even with expert knowledge. Therefore, the development of techniques for automatic hyperparameter tuning is highly desirable to improve the efficiency and accuracy of SJMs in practice. This aligns with recent trends in machine learning, where there is a substantial interest in automatic selection of hyperparameters, e.g. for the topology of deep neural networks (Hutter et al., 2019). Nystrup et al. (2020b) find that their model's performance improves significantly when they undertake hyperparameter tuning to optimize specific application-related performance criteria.

In general, researchers determine optimal hyperparameters using either cross-validation (CV, Stone, 1974), by empirically testing which parameters deliver the best out-of-sample performance, or employing an *information criterion* (IC) such as the *Akaike's information criterion* (AIC, Akaike, 1974) or the *Bayesian information criterion* (BIC, Schwarz, 1978). AIC typically selects the best predictor but may not accurately determine the correct model order. On the other hand, BIC asymptotically identifies the correct model order, but often results in poorer predictive performance.

In recent work on *generalized information criteria* (GIC), Fan and Tang (2013) consider the case where the number of features is comparable to, or substantially larger than, the number of observations. They demonstrate the inadequacy of AIC and BIC in high-dimensional settings for identifying the correct model. To address this issue, they introduce a novel IC, here referred to as the *Fan-Tang information criterion* (FTIC), which is proven to consistently select the correct model order asymptotically.

In the present Chapter, we adapt the GIC framework for model selection in high-dimensional penalized models to SJMs. In particular, we make three main contributions. First, to effectively leverage the GIC, we rephrase the SJM into an approximate log-likelihood framework, then we derive an expression that quantifies its model complexity, and we further extend the GIC framework to incorporate SJMs. Specifically, we propose a FTIC, an AIC, and a BIC for SJMs, each incorporating the proposed model complexity measure that penalizes for the number of active features, states, and jumps between states. Second, in a comprehensive simulation study, we demonstrate that the new FTIC, suitably modified for SJMs, outperforms the other ICs in selecting the correct hyperparameter values. Third, we conduct an empirical study where we apply a SJM to a dataset consisting of features related to global equity markets, as represented by the MSCI developed and emerging market indexes. We determine the best SJM based on the new FTIC, allowing

us to correctly detect switches between different phases of the market, and to identify its main drivers, which primarily hinges on features related to volatility.

The Chapter is organized as follows. In Section 4.2, we review some popular ICs for model selection and we propose an extension of the GIC framework for SJMs. In Section 4.3, we present a simulation study comparing the new FTIC, AIC, and BIC in selecting the correct model. In Section 4.4, we employ our model selection framework for SJMs to determine the number of regimes, their level of persistence, and the main drivers of the world equity markets. In Section 4.5, we draw some conclusions. In Appendix 4.A, we derive the approximate log-likelihood function for SJMs, and Appendix 4.B offers technical background on the measures used to evaluate the suitability of the proposed ICs.

4.2 Methodology

A fundamental result in theoretical statistics is the importance of the (log-)likelihood function, as it effectively encodes model specific information generated from observations. As such, it is often used for parameter inference.

However, it cannot be utilized for comparisons between competing models, since it increases monotonically with increasing model complexity. It is known from likelihood ratio tests (Wilks, 1938) that, due to overfitting, adding irrelevant parameters still increases the value of the log-likelihood function. Instead, researchers often use CV or ICs for model selection and hyperparameter estimation.

The CV technique involves partitioning the data into two or more blocks, using some for model fitting and some others for validation. This practice breaks the positive dependence between the parameter estimates and the model fit, as these are based on separate random events. Thus, the model fit computed on the validation set is not inflated by overfitting, but rather penalized by it.

To make the best possible use of the data, one typically performs random resampling of the blocks and averages the final result across each random experiment. A special case of CV is the leave-one-out CV: Stone (1977) shows that this methodology is asymptotically equivalent to the AIC, a property that holds for any model. Another special case is the leave-ν-out CV when $\nu = T(1 - 1/(\log(T) - 1))$, where T is the sample size. This approach is asymptotically equivalent to BIC for linear models (Shao, 1997).

However, although CV is easily understood in principle, it is also associated with a number of difficulties, such as computational cost and the difficulty to apply it to dependent data. Hence, ICs are still relevant to select model order and hyperparameters.

4.2.1 Information criteria

ICs provide a quantitative way to balance the in-sample goodness of fit of a model with its model complexity, thus avoiding overfitting. An IC is defined, in general, as a combination of two terms

$$IC := F + a_T M, \tag{4.1}$$

where F is a measure of model fit, often the log-likelihood function, a_T is a positive sequence depending on the sample size T, and possibly the number of parameters and/or features considered, and M is a measure of model complexity (Konishi and Kitagawa, 2008; Fan and Tang, 2013).

4.2.1.1 Akaike's information criterion

AIC is used for selecting the best predictive model. It is an asymptotic unbiased estimator of the expected Kullback-Leibler risk loss under the assumption that the candidate model includes the true model. It is given by

$$\text{AIC} := -2\ell(\hat{\boldsymbol{\theta}}) + 2q,$$

with q, the number of parameters, considered as measure of model complexity, $a_T = 2$, and employing minus twice the log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ as measure of model fitting. The closely related *Takeuchi information criterion* (TIC), introduced by Takeuchi (1976), generalizes AIC by accommodating likelihood misspecification.

4.2.1.2 Bayesian information criterion

Schwarz (1978) derives BIC by considering the posterior probability for a specific model, given the data. This is a consistent criteria, as it will asymptotically select the correct model with probability one. It is obtained by noting that the posterior probability for a model $\mathcal{M}$, given data $\boldsymbol{Y}$, is defined as

$$\begin{aligned}\mathbb{P}(\mathcal{M}|\boldsymbol{Y}) &\propto \mathbb{P}(\boldsymbol{Y}|\mathcal{M})\mathbb{P}(\mathcal{M}),\\ &= \int \left(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta},\mathcal{M})\mathbb{P}(\boldsymbol{\theta}|\mathcal{M})\right) \mathrm{d}\boldsymbol{\theta}\ \mathbb{P}(\mathcal{M}),\end{aligned}$$

where $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta},\mathcal{M})$ is the likelihood, $\mathbb{P}(\boldsymbol{\theta}|\mathcal{M})$ is the prior for the parameters, and $\mathbb{P}(\mathcal{M})$ is the prior distribution on the model space. Assuming uninformative priors and approximating the integral using the Laplace method results in

$$\text{BIC} := -2\ell(\hat{\boldsymbol{\theta}}) + q\log(T),$$

in accordance with Equation (4.1), with a_T set as $\log(T)$, and employing the same model fitting criteria and complexity measures as employed by AIC.

Previous research has extended the BIC concept to situations where the number of parameters q is of the same magnitude, or even substantially larger than the number of observations T. This setup causes some difficulties as the number of possible models grows rapidly with q. Chen and Chen (2008) addresses this issue by modifying the prior probabilities $\mathbb{P}(\mathcal{M})$, penalizing models with additional parameters. They show that their generalized BIC is still consistent for large models spaces.

4.2.2 Generalized information criteria

Fan and Tang (2013) present a further extension as they consider the setup where the number of features grows substantially faster than the number of observations, extending the Chen and Chen (2008) methodology. They demonstrate that, in high-dimensional setting, AIC or BIC fail to identify the correct model, whereas their criterion remains consistent.

Their article defines a GIC, which compares a saturated model and a candidate model, as

$$\text{GIC} := \frac{1}{T}\left\{2\left(\ell^S(\boldsymbol{Y}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}, \boldsymbol{Y})\right) + a_T M\right\}, \tag{4.2}$$

where $\ell^S(\boldsymbol{Y})$ is the log-likelihood for the saturated model, evaluated using the estimate of the complete, unconstrained parameter vector, while $\ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}, \boldsymbol{Y})$ is the log-likelihood evaluated using the parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$. Here, $\boldsymbol{\alpha}$ represents the set of active parameters, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ is the estimated parameter vector that exclusively pertains to the active set.

Fan and Tang (2013) suggest setting the model complexity M equal to the cardinality of the active set of parameters, and $a_T = \log(\log(T))\log(p)$, where p is the total number of features considered, cf. Hannan and Quinn (1979); Chen et al. (2009).

4.2.3 GIC for sparse statistical jump models

We adapt the GIC as given in Equation (4.2) to the statistical JMs and SJMs presented in Section 3.2 of Chapter 3. To the best of our knowledge, there is no exact closed form expression for the log-likelihood function, but we derive an approximation in Appendix 4.A, given by

$$\begin{aligned}\hat{\ell}(\boldsymbol{\mu}) &= \sum_{t=1}^{T}\left\{C(K,p) - \frac{1}{2}\|\widetilde{\boldsymbol{y}}_{t,p} - \boldsymbol{\mu}_{s_t}\|_2^2\right\} \\ &= T \cdot C(K,p) - \frac{1}{2}\sum_{t=1}^{T}\|\widetilde{\boldsymbol{y}}_{t,p} - \boldsymbol{\mu}_{s_t}\|_2^2,\end{aligned}$$

being $C(K,p) = -\log(K) - \frac{p}{2}\log(2\pi)$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ the conditional mean vectors of the features, and $\widetilde{\boldsymbol{y}}_{t,p}$ the standardized features. From Equation, (3.3), TSS = WCSS + BCSS, so it follows that the approximate log-likelihood can be expressed as

$$\hat{\ell}(\boldsymbol{\mu}) = T \cdot C(K,p) - \frac{1}{2}\text{WCSS} = T \cdot C(K,p) - \frac{1}{2}(\text{TSS} - \text{BCSS}),$$

with BCSS being computed in Equation (3.4). Finally, we obtain the approximation of the GIC by inserting those approximations in Equation (4.2)

$$\begin{aligned}\widehat{\text{GIC}} &= \frac{1}{T}\left\{2\left(\hat{\ell}^S(\boldsymbol{Y}) - \hat{\ell}(\hat{\theta}_{\boldsymbol{\alpha}}, \boldsymbol{Y})\right) + a_T M\right\} \\ &= \frac{1}{T}\left\{(\text{BCSS}^S - \text{BCSS}) + a_T M\right\} + 2\left(C^S(K,p) - C(K,p)\right), \end{aligned} \tag{4.3}$$

where $\boldsymbol{Y}$ is the matrix of standardized features.

4.2.3.1 Model complexity

Model complexity, denoted as M, is typically defined by the size of the active parameter set. This definition requires adaptation in the context of the SJM formulation, where three hyperparameters serve distinct roles: K governs the number of states, λ regulates the number of jumps, and κ controls feature selection.

The number of parameters for a Gaussian mixture model with equal probability for each component is $K|\boldsymbol{\alpha}|$, the number of states multiplied by the number of active features in each state. Additionally, the number of jumps (and the number of potential state transitions) will have a positive impact on the model complexity. This suggests a model complexity approximately given by

$$M \approx K(|\boldsymbol{\alpha}_\kappa| + \Gamma_\lambda), \tag{4.4}$$

where $|\boldsymbol{\alpha}_\kappa|$ is the number of features selected by the SJM, and $\Gamma_\lambda = \sum_t \mathbb{I}_{s_t \neq s_{t-1}}$ the number of jumps across states. Γ_λ is binomially distributed, with the number of jumps being the sufficient statistic for a binomial distribution.

We define the model complexity measure M as linear approximation of Equation (4.4)

$$M := K\left(|\boldsymbol{\alpha}_0| + \Gamma_0\right) + K_0\left(|\boldsymbol{\alpha}_\kappa| - |\boldsymbol{\alpha}_0| + \Gamma_\lambda - \Gamma_0\right), \tag{4.5}$$

near the point $(K_0, |\boldsymbol{\alpha}_0|, \Gamma_0)$. Equation (4.5) penalizes for increasing values of K, $|\boldsymbol{\alpha}_\kappa|$, and Γ_λ, the number of latent states, active features and jumps. $|\boldsymbol{\alpha}_\kappa|$ and Γ_λ depends indirectly on the hyperparameters κ and λ, respectively; the first increases with increasing values of κ, the latter decreases when increasing λ.

In practical applications, we recommend choosing K_0, $|\boldsymbol{\alpha}_0|$, and Γ_0 using pre-existing

insights about the number of latent states, features and jumps. By incorporating this prior knowledge into the linear approximation, we provide a solid basis for tailoring the model fit measure to our specific expectations.

To enhance the efficacy of the GIC, we set an upper limit on Γ_λ equal to 40% of the total number of observations, leveraging a heuristic approach to ensure its robustness. This serves the purpose of filtering out models characterized by excessively high number of jumps, in essence pure mixture models, which typically lack meaningful or insightful interpretations.

We determine hyperparameters $\bar{\lambda}$, $\bar{\kappa}$, and $\bar{K}$ for fitting the saturated model in the SJM framework as follows. Setting $\bar{\lambda} = 0$ and $\bar{\kappa} = \sqrt{p}$ is an easy choice, as these values result in an SJM with no jump penalty and that considers all the features. However, selecting $\bar{K}$ is not as clear. In our experience, if the goal is to estimate a model with recurrent states, we recommend not exceeding $\bar{K} = 6$ although that number is model and data dependent. When $\bar{K}$ is too high, the modified GIC selects a large number of states, each one being visited only once, which may not be desirable for the application in mind.

We can easily recover SJM modified versions of FTIC, AIC, and BIC based on (4.3) just changing the value of a_T. With a slight abuse of notation, we denote the resulting indexes as FTIC, AIC, and BIC.

4.3 Simulation study

We conduct two simulation studies to demonstrate the efficacy of the proposed FTIC, AIC, and BIC for SJMs. The studies aim to assess the ICs accuracy in correctly determining hyperparameter values, specifically the persistence parameter λ, the sparsity parameter κ, and the number of latent states K.

4.3.1 Simulation setup

4.3.1.1 Model

We simulate observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, from a multivariate Gaussian HMM with K_{true} latent states,

$$\boldsymbol{y}_t | s_t \sim \mathcal{N}(\boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}).$$

The latent process $\{s_t\}$ is a Markov chain of first order with K_{true} states, with initial probabilities $\pi_i = 1/K_{\text{true}}$, $\forall i = 1, \ldots, K_{\text{true}}$, and transition probability matrix denoted by $\boldsymbol{\Pi} \in \mathbb{R}^{K_{\text{true}} \times K_{\text{true}}}$, with elements π_{ij}, $i, j = 1, \ldots, K_{\text{true}}$. We vary the number of latent states $K_{\text{true}} \in \{2, 3, 4\}$ and simulate data in both studies. We draw state-conditional mean vectors $\boldsymbol{\mu}_k \in \mathbb{R}^{p_{\text{true}}}$, $k = 1, \ldots, K_{\text{true}}$, from the uniform distribution $\mathcal{U}(-2, 2)$. The state-conditional covariance matrices $\boldsymbol{\Sigma}_k$ have diagonal elements equal to 1 and off-diagonal elements $\rho_{ij}^{(k)}$,

$i, j = 1, \ldots, p_{\text{true}}$, $i \neq j$, $k = 1, \ldots, K_{\text{true}}$, given by:

- $\rho_{ij}^{(1)} = 0.80$, $\rho_{ij}^{(2)} = 0.40$ when $K_{\text{true}} = 2$;
- $\rho_{ij}^{(1)} = 0.80$, $\rho_{ij}^{(2)} = 0.60$, $\rho_{ij}^{(3)} = 0.30$ when $K_{\text{true}} = 3$;
- $\rho_{ij}^{(1)} = 0.80$, $\rho_{ij}^{(2)} = 0.60$, $\rho_{ij}^{(3)} = 0.30$, $\rho_{ij}^{(4)} = 0$ when $K_{\text{true}} = 4$.

4.3.1.2 Features construction

We derive the features employed in the simulation study through standardization of the simulated time-series $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$. Following Nystrup et al. (2021), to evaluate the SJM ability to select features, we consider two datasets. The first dataset includes only the original p_{true} features, while the second dataset contains a total of $p > p_{\text{true}}$ features. Within this set, the first p_{true} correspond to the original time-series, while we obtain the subsequent $p - p_{\text{true}}$ through distinct permutations of the rows of the original data matrix. The purpose of these row permutations is to disrupt temporal information while preserving both cross-sectional details and distributional attributes. We refer to these as *false* features. Specifically, each simulation study incorporates $p = 300$ features in total, $p - p_{\text{true}} = 200$ of which are false features.

4.3.1.3 Estimation

We fit a family of SJMs by varying λ, κ and K and we compute the three ICs for each possible λ, κ, and K according to Equation (4.3). To determine M, we take K_0 and $|\boldsymbol{\alpha}_0|$ to be equal to their true counterparts, i.e. $|\boldsymbol{\alpha}_0| = 100$ and $K_0 = K_{\text{true}}$. We set Γ_0 equal to $(1 - \pi_{ii})(K_0 - 1)T$ based on our empirical observation that this value closely approximates the average true number of jumps.

In the first study, we vary the number of observations considering three possible setups, each one consisting of 300, 600, and 1,000 observations, respectively. The objective here is to evaluate the convergence of the ICs. We emphasize that the application employs a substantially larger number of observations.

In the second study, we investigate model performance across varying degrees of persistence. We achive this by exploring different values of the self-transition probabilities π_{ii}, $i = 1, \ldots, K_{\text{true}}$, representing the probability of remaining within a specific latent state. In particular, we consider a *less persistent* HMM with $\pi_{ii} = 0.70$, and a *more persistent* HMM with $\pi_{ii} = 0.90$.

4.3.1.4 Evaluation of the performance

We repeat the procedure 100 times for each scenario, each time changing the seed, and the reported results refer to the average indexes obtained across the 100 studies. To provide a better understanding of the results, we compute the confidence bands for the 100 simulations as the average ICs ± 2 the standard deviation. These bands enable us to assess the variability of the ICs across the different simulations.

To evaluate the ability of the ICs in selecting optimal λ and κ values, we employ the adjusted Rand index (ARI, Hubert and Arabie, 1985), which measures the similarity between true and predicted cluster assignments. We calculate ARI for both the estimated state sequences $\{\hat{s}_t\}$ and active feature sequences $\{\omega_i\}$. We obtain $\{\omega_i\}$ through the following

$$\omega_i = \mathbb{I}_{\hat{w}_i \neq 0}, \quad i = 1, \ldots, p,$$

where $\hat{w}_i$ is the estimated weight for feature i. We compare this sequence to that of true active features, given by

$$\underbrace{(1, 1, \ldots, 1,}_{p_{\text{true}} \text{ times}} \underbrace{0, 0, \ldots, 0)}_{p - p_{\text{true}} \text{ times}},$$

which is a sequence with 1s in the first p_{true} entries and 0 elsewhere. We denote the two ARIs by ARI($\{\hat{s}_t\}$) and ARI($\{\omega_i\}$), respectively. We recall that an ARI equal to one corresponds to a perfect match between the elements of the two sequences, and the index decreases as similarity decreases. For an explanation of how to calculate this index, please refer to Appendix 4.B.

4.3.2 Varying the number of observations

In this initial scenario, we consider a dataset denoted by $\boldsymbol{Y} \in \mathbb{R}^{T \times p}$, with $p = 300$, consisting of 100 true and 200 false features. Additionally, we explore three distinct values for T, namely 300, 600, and 1,000. We set the transition probability matrices as follows

$$\boldsymbol{\Pi}_2 = \begin{bmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{bmatrix},$$

$$\boldsymbol{\Pi}_3 = \begin{bmatrix} 0.80 & 0.10 & 0.10 \\ 0.10 & 0.80 & 0.10 \\ 0.10 & 0.10 & 0.80 \end{bmatrix},$$

$$\mathbf{\Pi}_4 = \begin{bmatrix} 0.80 & 0.0\bar{6} & 0.0\bar{6} & 0.0\bar{6} \\ 0.0\bar{6} & 0.80 & 0.0\bar{6} & 0.0\bar{6} \\ 0.0\bar{6} & 0.0\bar{6} & 0.80 & 0.0\bar{6} \\ 0.0\bar{6} & 0.0\bar{6} & 0.0\bar{6} & 0.80 \end{bmatrix},$$

for the HMMs with 2, 3 and 4 states, respectively. We fit a family of SJMs by varying λ, κ, and K such that $\lambda \in \{0, 5, 10, 25, 50, 100\}$, $\kappa \in \{1, 2, 3, \ldots, 10, 12, 14, 17\}$, and $K \in \{2, 3, 4\}$.

Table 4.1 presents the results of the GIC and two ARIs computed for different values of K_{true} and T. Additionally, it presents the average count of correctly identified features (True Positives, TP) and incorrectly identified features (False Positives, FP). Based on these findings, it can be concluded that FTIC proves effective in selecting the three hyperparameters. In fact, it fails to select the correct number of latent states only when $T = 300$ and $K_{\text{true}} = 4$. BIC exhibits shortcomings in correctly selecting the number of states when $K_{\text{true}} = 2$ and T is either 300 or 600, whereas it performs good when $T = 1,000$. Conversely, AIC always fails in selecting the appropriate number of latent states when the true count is 2. When it comes to selecting the jump penalty λ and the sparsity hyperparameter κ, FTIC generally outperforms both AIC and BIC. The minimum FTIC values correspond to a λ value that yields $\text{ARI}(\{s_t\}) = 1$ in all scenarios except one ($T = 300$ and $K_{\text{true}} = 4$), and the minimum FTIC corresponds to a κ value that maximizes $\text{ARI}(\{\omega_i\})$ in most scenarios, followed by BIC. Regarding the selection of features, AIC demonstrates a tendency to include a considerable number of false features, even in scenarios with a high number of observations. This aligns with a common observation in regression analysis, where, with larger sample sizes, AIC tends to include an increasing number of irrelevant explanatory variables (Heinze et al., 2018). In contrast, FTIC adopts a more conservative approach, selecting a substantial portion of the true features across most scenarios.

Figures 4.1, 4.2, and 4.3 display the FTIC, AIC, and BIC values for different true numbers of latent states when T is set to 300, 600, and 1,000, respectively. For each plot, we vary κ (λ) within the pre-specified interval, while the value of λ (κ) is selected as the value that minimizes the corresponding index. The results indicate that AIC and BIC typically fails to identify the correct number of latent states when $K_{\text{true}} = 2$. In contrast, FTIC performs well in almost all scenarios, showing decreasing variability as the number of states increases. This indicates that FTIC is more precise in estimating the true number of states compared to the other two ICs.

Based on these results, we can conclude that FTIC is preferable to the other two indexes since it yields satisfactory results for all three hyperparameters selections, even when the number of observations is low.

4.3.3 Varying the self-transition probability

In this second scenario, we assess the effectiveness of the proposed ICs for varying levels of persistence within a given state. We consider a dataset $\boldsymbol{Y}$ consisting of $T = 1,000$ observations and $p = 300$ features, of which 200 are false features. We set the matrices of transition probabilities equal to

$$\boldsymbol{\Pi}_2 = \begin{bmatrix} 0.70 & 0.30 \\ 0.30 & 0.70 \end{bmatrix}, \tag{4.6}$$

$$\boldsymbol{\Pi}_3 = \begin{bmatrix} 0.70 & 0.15 & 0.15 \\ 0.15 & 0.70 & 0.15 \\ 0.15 & 0.15 & 0.70 \end{bmatrix}, \tag{4.7}$$

$$\boldsymbol{\Pi}_4 = \begin{bmatrix} 0.70 & 0.10 & 0.10 & 0.10 \\ 0.10 & 0.70 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.70 & 0.10 \\ 0.10 & 0.10 & 0.10 & 0.70 \end{bmatrix}, \tag{4.8}$$

for the less persistent HMMs with 2, 3 and 4 states, respectively. The transition matrices are equal to

$$\boldsymbol{\Pi}_2 = \begin{bmatrix} 0.90 & 0.10 \\ 0.10 & 0.90 \end{bmatrix}, \tag{4.9}$$

$$\boldsymbol{\Pi}_3 = \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{bmatrix}, \tag{4.10}$$

$$\boldsymbol{\Pi}_4 = \begin{bmatrix} 0.90 & 0.0\bar{3} & 0.0\bar{3} & 0.0\bar{3} \\ 0.0\bar{3} & 0.90 & 0.0\bar{3} & 0.0\bar{3} \\ 0.0\bar{3} & 0.0\bar{3} & 0.90 & 0.0\bar{3} \\ 0.0\bar{3} & 0.0\bar{3} & 0.0\bar{3} & 0.90 \end{bmatrix}, \tag{4.11}$$

for the more persistent HMMs with 2, 3 and 4 states, respectively. We fit a sequence of SJMs by varying λ, κ and K within the same parameter sets as the first simulation study. We then compute the ICs for each possible combination of λ, κ, and K.

In Table 4.2, we present the results for this second simulation study, which reports the IC values and the corresponding ARIs in the less persistent and in the more persistent scenarios. Overall, the results show that FTIC outperforms the other two ICs, as it always selects the correct number of latent states and produces good ARI results, except in the less persistent scenario when $K_{\text{true}} = 4$. Nevertheless, we find this result satisfying as

the objective of the study is to test the limits of the parameter settings, and assess the performance of the proposed ICs. In fact, it is worth noting that, in practical applications, the estimated persistence is typically around 90% (Ang and Timmermann, 2012).

FTIC is conservative in terms of the average number of selected features, as it includes the majority of true features while incorporating only a minimal proportion, or even none at all, of false ones. AIC and BIC face challenges when $K_{\text{true}} = 2$ in the less persistent scenario. An important observation is that all ICs perform good as we increase the level of persistence.

Figures 4.4 and 4.5 show the average estimated values of FTIC, AIC, and BIC over 100 simulation samples, as well as the corresponding confidence bands, for varying values of π_{ii}. Once again, FTIC usually achieves the minimum value when the true number of latent states is selected. However, in the less persistent scenario, AIC and BIC face difficulties in accurately determining the number of latent states when $K_{\text{true}} = 2$. The results for AIC are less clear due to overlapping confidence bands, making it challenging to draw definitive conclusions.

In summary, we can conclude that FTIC outperforms other indexes in all scenarios, even when the number of observations slightly exceed the number of features. For hyperparameter selection in the SJM framework, FTIC stands as the best choice, while BIC proves valuable when the sample size is high. The simulation results also confirm the reliability of the first-order approximation used in IC derivation.

4.4 An application to the MSCI and MSCIEM indexes

Regime switching models are commonly used to determine the state of financial markets and have proven valuable in asset allocation and risk management (Ang and Timmermann, 2012; Nystrup et al., 2015, 2017, 2019; Yao et al., 2020). In this application, we use a set of features based on daily dollar-denominated log-returns and trading volumes of the MSCI World Index (MSCI) and MSCI Emerging Market Index (MSCIEM) to infer the states of the global equity market. The MSCI and MSCIEM represent large and mid-cap stocks across 23 developed markets and 24 emerging markets countries, respectively.[2,3] Each index covers approximately 85% of the free float-adjusted market capitalization in each country.

[2]The developed markets countries are: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Hong Kong, Ireland, Israel, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the UK and the US. The MSCI World Index fact sheet is available at: https://www.msci.com/documents/10199/178e6643-6ae6-47b9-82be-e1fc565ededb.

[3]The emerging markets countries are: Brazil, Chile, China, Colombia, Czech Republic, Egypt, Greece, Hungary, India, Indonesia, Korea, Kuwait, Malaysia, Mexico, Peru, Philippines, Poland, Qatar, Saudi Arabia, South Africa, Taiwan, Thailand, Turkey and United Arab Emirates. The MSCI Emerging Markets fact sheet is available at: https://www.msci.com/documents/10199/c0db0a48-01f2-4ba9-ad01-226fd5678111.

With a slight abuse of notation, in the following we refer to the logarithmic difference of a time-series as its *log-return* even if the time-series is not a tradable asset or security.

We compute exponentially weighted moving averages (EMA) of log-returns (r), volatilities (σ) and log-returns of volumes (V) with half-lives equal to 1, 2, 7, 14, 30 and 90 days. We include exponentially weighted linear (ρ) and Gerber (g) (Gerber et al., 2022) correlations between MSCI and MSCIEM log-returns and between log-returns and log-returns of volumes for each index, with the same half-lives as above.

We add the following momentum-based features: the time-series momentum signal (RF) of Moskowitz et al. (2012) with 1, 2, 7, 14, 30 and 90 daily lags; the relative strength index (RSI) with a lag of 14, and the moving average convergence divergence minus signal (MACDS) indicator, computed by subtracting the 26-period EMA from the 12-period EMA and further subtracting the 9-period EMA (known as the "signal line"). Additionally, we consider the Amihud (2002) illiquidity measure (AMIHUD) calculated as the ratio between absolute log-returns and volumes for each index. Finally, we consider log-return of the VIX index along with its EMAs with the same half-lives as above. We also utilize exponential weighted linear and Gerber correlations between the VIX log-returns and log-returns for MSCI and MSCIEM, respectively.

The final dataset consists of $p = 125$ features spanning for the time period[4] January 30th, 1996 - March 3rd 2023, for a total number of observations $T = 6,843$. We fit a sequence of SJMs for varying λ, κ, and K, then we select the best model according to the FTIC, computed for $K_0 = 3$, $|\boldsymbol{\alpha}_0| = 50$ and $\Gamma_0 = (1 - \pi_{ii})(K_0 - 1)T$, with $\pi_{ii} = 0.875$. Our choice is justified by the expectation of the practical interpretation, with an anticipated number of states likely to be either 2 or 3 (Guo et al., 2011; Ang and Timmermann, 2012; Dua and Tuteja, 2021). Regarding the selected features, we anticipate around 50, encompassing those related to volatility and cross-correlation, which aligns with prior studies (Nystrup et al., 2020a, 2021). It is worth noting that, while the model selection procedure remains robust to the choice of K_0 and $|\boldsymbol{\alpha}_0|$, it is sensitive to Γ_0. Therefore, we recommend setting this value sufficiently high, as otherwise the FTIC may opt for a higher number of states.

Results show that the minimum FTIC corresponds to $\lambda = 10$, $\kappa = 6$, and $K = 3$. The daily state-conditional means and volatilities of MSCI log-returns are -0.08%, 0.00%, and 0.04%, and 2.57%, 1.00%, and 0.68%. Daily state-conditional means and volatilities of MSCIEM log-returns are -0.05%, -0.02%, and 0.04%, and 2.73%, 1.20%, and 0.82%, respectively. Moreover, the daily state-conditional pairwise correlations are 77%, 66% and 45%. Based on these findings, we can characterize each state as bear, neutral, and bull

[4]The MSCIEM made its debut on January 1, 2001, and data preceding this date is estimated based on the index's hypothetical performance during that earlier period. Please refer to the index documentation for more details https://www.msci.com/documents/10199/c0db0a48-01f2-4ba9-ad01-226fd5678111.

market regimes, respectively.

The sojourn times for each state are 87, 114, and 122.65 days, and the SJM spends 5.09%, 48.31%, and 46.60% of its time in each of them. Figure 4.6 depicts the cumulative log-returns of the two indexes with the estimated states highlighted with different colours. We underline that the model is remarkably able to track rise/drops in cumulative returns, as increasing returns are always highlighted in green (bull state) and decreasing returns in yellow or red (neutral and bear states).

We now examine the features selected by the SJM. Out of the original 125, 55 have been identified as relevant, as reported in Table 4.3. EMAs for volatilities possess the highest weights, and their state-dependent values align perfectly with the previously defined regime characterization. EMAs for MSCI and MSCIEM log-returns with half-lives of 90 and 30 are also relevant variables, and their state-dependent values consistently decrease from the bull to the bear state. The correlations between the log-returns of MSCI and MSCIEM demonstrate an upward trend from the bull to the bear states. This fact is commonly observed during periods of financial crisis, where the correlation among assets of the same type tends to strengthen. Additionally, the correlations between the log-returns of the VIX index and the log-returns of MSCI and MSCIEM are significant and consistently exhibit a negative trend, declining in magnitude from bear to bull states. This pattern indicates that an increase in market volatility corresponds to negative log-returns, as anticipated.

Table 4.4 shows feature weights categorized into groups: volatility, correlation of MSCI and MSCIEM log-returns with VIX, momentum, and cross-correlation. The dominant role of volatility in explaining equity markets is evident, with correlation with VIX also contributing significantly. Momentum-related features hold a total weight of 8.01%, and cross-correlation, specifically between MSCI and MSCIEM, accounts for 19.96%.

4.5 Discussion

We propose a modified version of the generalized information criteria to perform hyperparameter selection in the sparse statistical jump model framework. This involves rephrasing the sparse statistical jump model into an approximate log-likelihood framework, followed by the derivation of an expression to quantify its model complexity. Through two simulation studies, we test the ability of FTIC, AIC, and BIC suitably modified for sparse statistical jump models to correctly select the hyperparameter values. In the first simulation study, varying the number of observations, we show that FTIC outperforms the other two information criteria in correctly identifying the true number of latent states, number of relevant features, and level of persistence within a specific state. In fact, it provides better results in terms of the adjusted Rand index and it always selects the correct number of states and features, except when the true number of latent states is 2 and the number of observations

is low. BIC provides good results for sufficiently high number of observations, while AIC tends to include a considerable number of false features. In the second simulation study, we evaluate the performance of the proposed information criteria across different levels of persistence. The results indicate that all information criteria deliver excellent results when the persistence is high. However, in scenarios with low persistence, FTIC outperforms the other two information criteria due to its more conservative approach in selecting the number of states and features. Based on these observations, we can conclude that FTIC is the best information criterion for the sparse statistical jump model framework, given its better performance in selecting the hyperparameters, followed by BIC.

We present an application regarding the equity market and we show that the selected model is meaningful as it has a valid economic interpretation. In fact, each of the states represents a different market scenario, from bull to bear, the state-jumps are coherent with rising/falling price periods and the selected features have attractive financial explanations.

Appendices

4.A Approximate log-likelihood function for JMs

In this appendix, we derive an approximation for the log-likelihood function of statistical jump models. We assume that data, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, have been standardized with zero mean and unit variance.

For a sequence of distinct real numbers $v_1, \ldots, v_K \in \mathbb{R}$, with $v_i \neq v_j$, for all $i \neq j$, we denote by $v_{(1)}, \ldots, v_{(K)}$, the values of the original sequence sorted in ascending order, $v_{(1)} < \ldots < v_{(K)}$.

Lemma 1 (LogSumExp)**.** *Suppose $v_1, \ldots, v_K$, is a sequence of distinct real numbers. Then,*

$$\log\left(\sum_{k=1}^{K}\exp(v_k)\right) = v_{(K)} + \log\left(1 + \sum_{k=1}^{K-1}\exp\left(v_{(k)} - v_{(K)}\right)\right).$$

Proof. We observe that

$$\begin{aligned}
\log\left(\sum_{k=1}^{K}\exp(v_{(k)})\right) &= \log\left(\exp(v_{(K)})\sum_{k=1}^{K}\exp\left(v_{(k)} - v_{(K)}\right)\right) \\
&= v_{(K)} + \log\left(\sum_{k=1}^{K}\exp\left(v_{(k)} - v_{(K)}\right)\right) \\
&= v_{(K)} + \log\left(1 + \sum_{k=1}^{K-1}\exp\left(v_{(k)} - v_{(K)}\right)\right).
\end{aligned}$$

□

Gaussian mixture models

We consider the Gaussian mixture model (GMM, McLachlan and Peel, 2000) with probability density function given by

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \gamma_k \phi\left(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) ,$$

where the k-th component, $\phi\left(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$, is the Gaussian probability density function with mean $\boldsymbol{\mu}_k \in \mathbb{R}^p$, and covariance $\boldsymbol{\Sigma}_k \in \mathbb{R}_+^{p \times p}$. $\gamma_k \in \mathbb{R}_+$ is the weight for the k-th component, and $\sum_{k=1}^{K} \gamma_k = 1$. For the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, the resulting log-likelihood function is given by

$$\ell\left(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{t=1}^{T} \log\left(\sum_{k=1}^{K} \gamma_k \phi\left(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right) , \tag{4.12}$$

where $\boldsymbol{\gamma} := \{\gamma_1, \ldots, \gamma_K\}$, $\boldsymbol{\mu} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, and $\boldsymbol{\Sigma} := \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$.

Proposition 1. *The log-likelihood function* (4.12) *can be expressed as*

$$\ell\left(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{t=1}^{T} \left(v_{(K),t} + \log\left(1 + \sum_{k=1}^{K-1} \exp\left(v_{(k),t} - v_{(K),t}\right)\right)\right) ,$$

where

$$v_{(i),t} := C(\gamma_{(i)}, \boldsymbol{\Sigma}_{(i)}) - \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_{(i)})^\top \boldsymbol{\Sigma}_{(i)}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_{(i)}) , \quad i = 1, \ldots K , \tag{4.13}$$

is an ascending sequence such that

$$v_{(1),t} < \ldots < v_{(K),t} , \quad t = 1, \ldots, T . \tag{4.14}$$

Furthermore, $C(\gamma_{(i)}, \boldsymbol{\Sigma}_{(i)}) \in \mathbb{R}$ *depends on* $\gamma_{(i)}$ *and* $\boldsymbol{\Sigma}_{(i)}$*, but not on* $\boldsymbol{y}_t$ *or* $\boldsymbol{\mu}_{(i)}$*.*

Proof. By defining

$$\exp(v_{k,t}) := \gamma_k \phi\left(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) , \tag{4.15}$$

we observe that

$$\begin{aligned} v_{k,t} &= \log(\gamma_k) + \log\left(\phi\left(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right) \\ &= \underbrace{\log(\gamma_k) - \frac{1}{2}\log\left(\det\left(2\pi\boldsymbol{\Sigma}_k\right)\right)}_{\text{Does not depend on } \boldsymbol{y}_t \text{ or } \boldsymbol{\mu}_k} - \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_k) \\ &= C(\gamma_k, \boldsymbol{\Sigma}_k) - \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)\,, \end{aligned}$$

where $C(\gamma_k, \boldsymbol{\Sigma}_k) := \log(\gamma_k) - \frac{1}{2}\log\left(\det\left(2\pi\boldsymbol{\Sigma}_k\right)\right)$. For every t, the sequence $\{v_{k,t}\}_{k=1}^K$ is made of distinct elements with probability one. The proof now follows immediately from Lemma 1. □

K-means

The K-means algorithm simplifies the GMM, yielding a computationally less complex problem. In particular, let us consider a GMM in which all components have the same covariance, and equal weights:

$$\boldsymbol{\Sigma}_k \equiv \epsilon \mathbf{I}_p \;\text{ and }\; \gamma_k \equiv 1/K\,, \quad k = 1, \ldots, K\,. \tag{4.16}$$

It is well-known that the K-means algorithm can be derived as the limit of the EM algorithm for this GMM, as $\epsilon \to 0$ (Bishop, 2006). It follows from Bishop (2006, Sec. 9.3.2), that the K-means algorithm finds the solution to

$$\hat{\boldsymbol{\mu}} = \arg\max \sum_{t=1}^{T} v_{(K),t} = \arg\min \sum_{t=1}^{T} \min_k \|\boldsymbol{y}_t - \boldsymbol{\mu}_k\|_2^2\,,$$

where $\hat{\boldsymbol{\mu}} := \{\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K\}$ are the K-means centroids. This expression is remarkably similar to the one in Nystrup et al. (2020c),

$$\hat{\boldsymbol{\mu}}^{K\text{-means}} = \arg\min \sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\mu}_{s_t}\|_2^2\,,$$

where $\{s_t\}$ is the latent state sequence. Below, we will use this result to derive an approximate log-likelihood function for JMs.

Proposition 2. *An approximate log-likelihood function for the K-means model is given by*

$$\ell\left(\boldsymbol{\gamma}, \boldsymbol{\mu}\right) = \sum_{t=1}^{T} v_{(K),t}\,,$$

where

$$v_{k,t} := C(\gamma_k, \boldsymbol{\Sigma}_k) - \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)\,, \quad i = 1, \dots K\,. \tag{4.17}$$

Furthermore, $C(\gamma_k, \boldsymbol{\Sigma}_k) \in \mathbb{R}$ *depends on* γ_k *and* $\boldsymbol{\Sigma}_k$*, but not on* $\boldsymbol{y}_t$ *or* $\boldsymbol{\mu}_k$*.*

Proof. By substituting (4.16) into formula (4.17), we obtain

$$\begin{aligned} v_{k,t} &= \log(\gamma_k) - \frac{1}{2}\log\left(\det\left(2\pi\boldsymbol{\Sigma}_k\right)\right) - \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_k) \\ &= C_\epsilon - \frac{1}{2\epsilon} d_{k,t}\,, \end{aligned}$$

where $C_\epsilon := C(1/K, \epsilon \mathbf{I}_p)$, for all k, and $d_{k,t} := \|\boldsymbol{y}_t - \boldsymbol{\mu}_k\|_2^2$, is the quadratic distance between observation $\boldsymbol{y}_t$, and centroid $\boldsymbol{\mu}_k$. By arranging the sequence $\{d_{k,t}\}_{k=1}^K$ in ascending order, $d_{(1),t} < \ldots < d_{(K),t}$, it follows that

$$v_{(K),t} = C_\epsilon - \frac{1}{2\epsilon} d_{(1),t}\,,$$

and

$$v_{(k),t} - v_{(K),t} = \frac{1}{2\epsilon}\left(d_{(1),t} - d_{(K-k+1),t}\right)\,, \quad k = 1, \dots, K-1\,.$$

Clearly, $d_{(1),t} - d_{(K-k+1),t} < 0$, for all $k < K$, which implies that $v_{(k),t} - v_{(K),t} \to -\infty$, and consequently, $\exp(v_{(k),t} - v_{(K),t}) \to 0$, as $\epsilon \downarrow 0$. Therefore,

$$\begin{aligned} \log\left(\sum_{k=1}^K \exp(v_{(k),t})\right) &= v_{(K),t} + \log\left(1 + \sum_{k=1}^{K-1} \exp\left(v_{(k),t} - v_{(K),t}\right)\right) \\ &\to v_{(K),t} \quad \text{as} \quad \epsilon \downarrow 0\,. \end{aligned}$$

□

Approximating the log-likelihood function

The SJM algorithm in Nystrup et al. (2021) closely resembles the K-means algorithm when computing the centroids $\hat{\boldsymbol{\mu}}^{K\text{-means}}$. The assumptions used in the K-means derivations, using Lemma 1, provide a data compression argument that reduces the sum across all K components in the mixture to only the most important single component. The corresponding approximation of the log-likelihood function is given by

$$\hat{\ell}\left(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{t=1}^T v_{(K),t}.$$

For the SJM, we recall that the data is *standardized*. Hence, we propose to use this first order approximation of the log-likelihood function, but evaluating it using $\gamma_k = 1/K$ and $\Sigma_k = \mathbf{I}_p$, arriving at

$$\begin{aligned}\hat{\ell}(\boldsymbol{\mu}) &= \sum_{t=1}^{T}\left(-\log(K) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\min_k \|\boldsymbol{y}_t - \boldsymbol{\mu}_k\|_2^2\right) \\ &= T\left(-\log(K) - \frac{p}{2}\log(2\pi)\right) - \sum_{t=1}^{T}\frac{1}{2}\min_k \|\boldsymbol{y}_t - \boldsymbol{\mu}_k\|_2^2.\end{aligned}$$

It should be clear that the argument maximizing this equation is the same as the K-means estimate of the centroids. However, the additional terms does have implications for the value of the log-likelihood function when varying the number of states, K, and the dimension, p.

4.B Adjusted Rand index

In this appendix, we provide some details on the adjusted Rand index (ARI, Hubert and Arabie, 1985). ARI measures the degree of overlapping between two grouping of partitions: $G^m = \{G_1^m, \ldots, G_h^m\}$, obtained with a clustering algorithm on a set of n elements, and $G^r = \{G_1^r, \ldots, G_l^r\}$, interpreted as the real clustering. Given the following contingency table

	G_1^m	G_2^m	$\cdots$	G_h^m	$\sum_{j=1}^{h} n_{ij}$
G_1^r	n_{11}	n_{12}	$\cdots$	n_{1h}	a_1
G_2^r	n_{21}	n_{22}	$\cdots$	n_{2h}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G_l^r	n_{l1}	n_{l2}	$\cdots$	n_{lh}	a_l
$\sum_{i=1}^{l} n_{ij}$	b_1	b_2	$\cdots$	b_h	n

$\mathrm{ARI}(G^r, G^m)$ is obtained as proposed in Das and Biswas (2023)

$$\mathrm{ARI}\left(G^r, G^m\right) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}.$$

Here, n_{ij} represents the count of shared elements within clusters G_i^m and G_j^r, while a_i aggregates all n_{ij} values linked to any G_j^r in G^r and all G_i^m in G^m. Similarly, b_j sums all n_{ij} values related to any G_i^m in G^m and all G_j^r in G^r.

Table 4.1: Simulation results for the minimum FTIC, AIC, and BIC and the corresponding λ, κ, and K for varying number of true latent states K_{true} when the number of observations T is equal to 300 (a) 600 (b) and 1,000 (c). Value refers to the averages across 100 simulations of the estimated value of the reported IC, and ARI($\{\hat{s}_t\}$) and ARI($\{\omega_i\}$) are the average ARIs computed between true and estimated sequences of states, and between true and estimated sequences of active features, respectively. TP (true positives) and FP (false positives) are the average numbers of correctly selected and wrongly selected features, respectively.

(a) $T = 300$

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_i\}$)	TP	FP
2	FTIC	44.08	5.00	7.00	2	1.00	0.55	63.38	0.00
	AIC	34.55	5.00	12.00	4	0.53	0.00	99.85	197.81
	BIC	44.72	5.00	12.00	4	0.53	0.00	99.85	197.81
3	FTIC	9.82	5.00	8.00	3	1.00	0.73	78.52	0.00
	AIC	15.38	5.00	9.00	3	1.00	0.74	97.66	22.48
	BIC	13.23	5.00	8.00	3	1.00	0.73	78.52	0.00
4	FTIC	-18.21	25.00	1.00	2	0.07	0.01	0.99	0.01
	AIC	-2.93	5.00	9.00	4	1.00	0.93	94.68	0.00
	BIC	-9.12	5.00	9.00	4	1.00	0.93	94.68	0.00

(b) $T = 600$

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_i\}$)	TP	FP
2	FTIC	43.55	5.00	7.00	2	1.00	0.55	63.43	0.00
	AIC	35.57	5.00	12.00	4	0.53	0.00	99.96	198.86
	BIC	43.00	5.00	12.00	4	0.53	0.00	99.96	198.86
3	FTIC	7.84	5.00	8.00	3	1.00	0.72	78.26	0.00
	AIC	12.69	5.00	9.00	3	1.00	0.78	97.88	22.16
	BIC	10.53	5.00	9.00	3	1.00	0.78	97.88	22.16
4	FTIC	-19.68	5.00	9.00	4	1.00	0.93	94.45	0.00
	AIC	-5.69	5.00	9.00	4	1.00	0.93	94.45	0.00
	BIC	-12.86	5.00	9.00	4	1.00	0.93	94.45	0.00

(c) $T = 1,000$

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_i\}$)	TP	FP
2	FTIC	43.14	5.00	7.00	2	1.00	0.55	63.33	0.00
	AIC	40.08	5.00	12.00	4	0.54	0.00	100.00	199.98
	BIC	43.44	5.00	7.00	2	1.00	0.55	63.33	0.00
3	FTIC	7.00	5.00	9.00	3	1.00	0.77	97.83	26.44
	AIC	11.72	5.00	9.00	3	1.00	0.77	97.83	26.44
	BIC	9.16	5.00	9.00	3	1.00	0.77	97.83	26.44
4	FTIC	-21.55	5.00	9.00	4	1.00	0.93	94.47	0.00
	AIC	-6.94	5.00	9.00	4	1.00	0.93	94.47	0.00
	BIC	-14.88	5.00	9.00	4	1.00	0.93	94.47	0.00

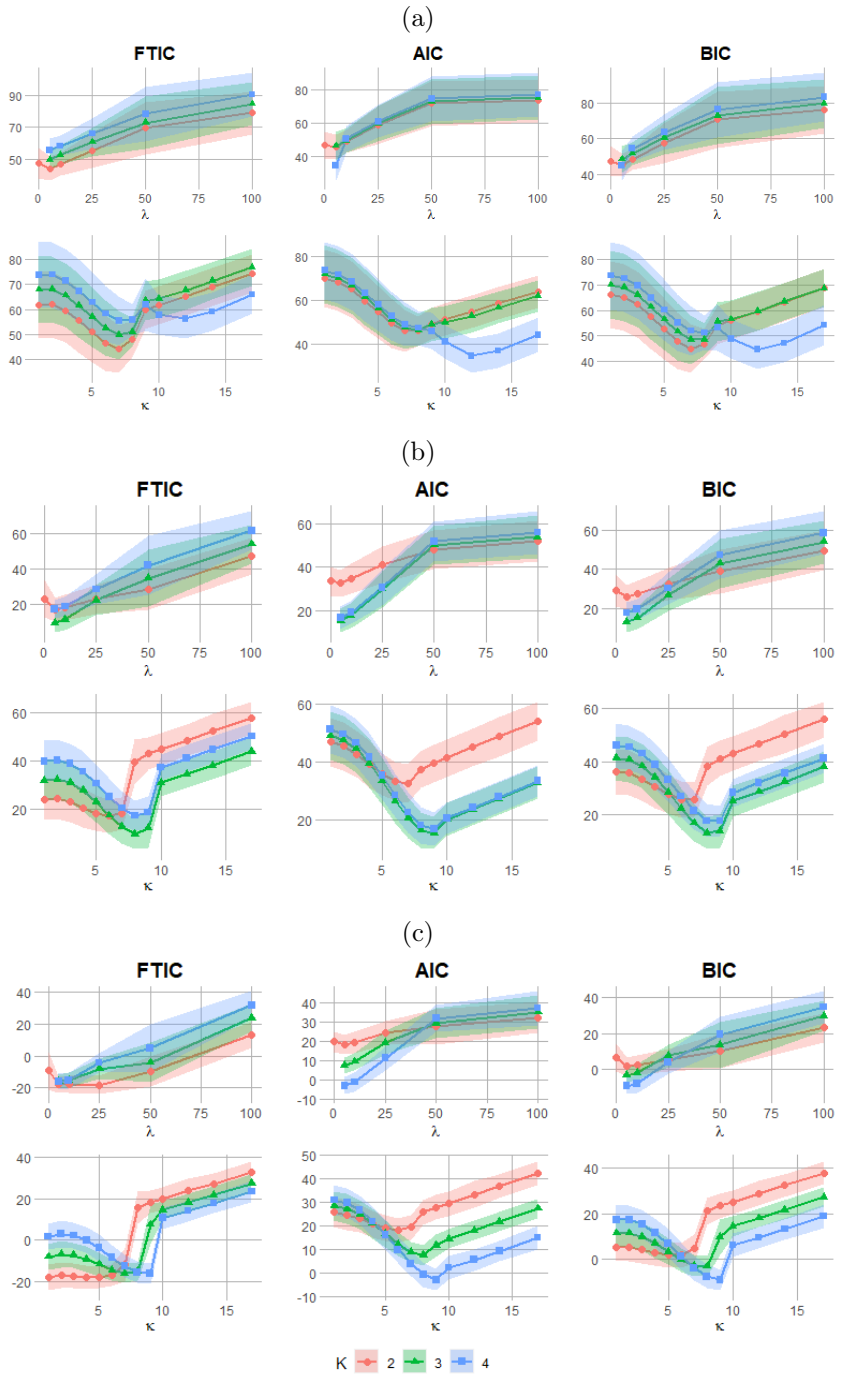


Figure 4.1: FTIC, AIC, and BIC of the SJM from 100 simulations each of length $T = 300$ for different values of λ and κ. Panel (a)–(c) depict the average ICs and their ± 2 standard deviation confidence bands when the true number of latent states (K_{true}) is equal to 2, 3, and 4, respectively.

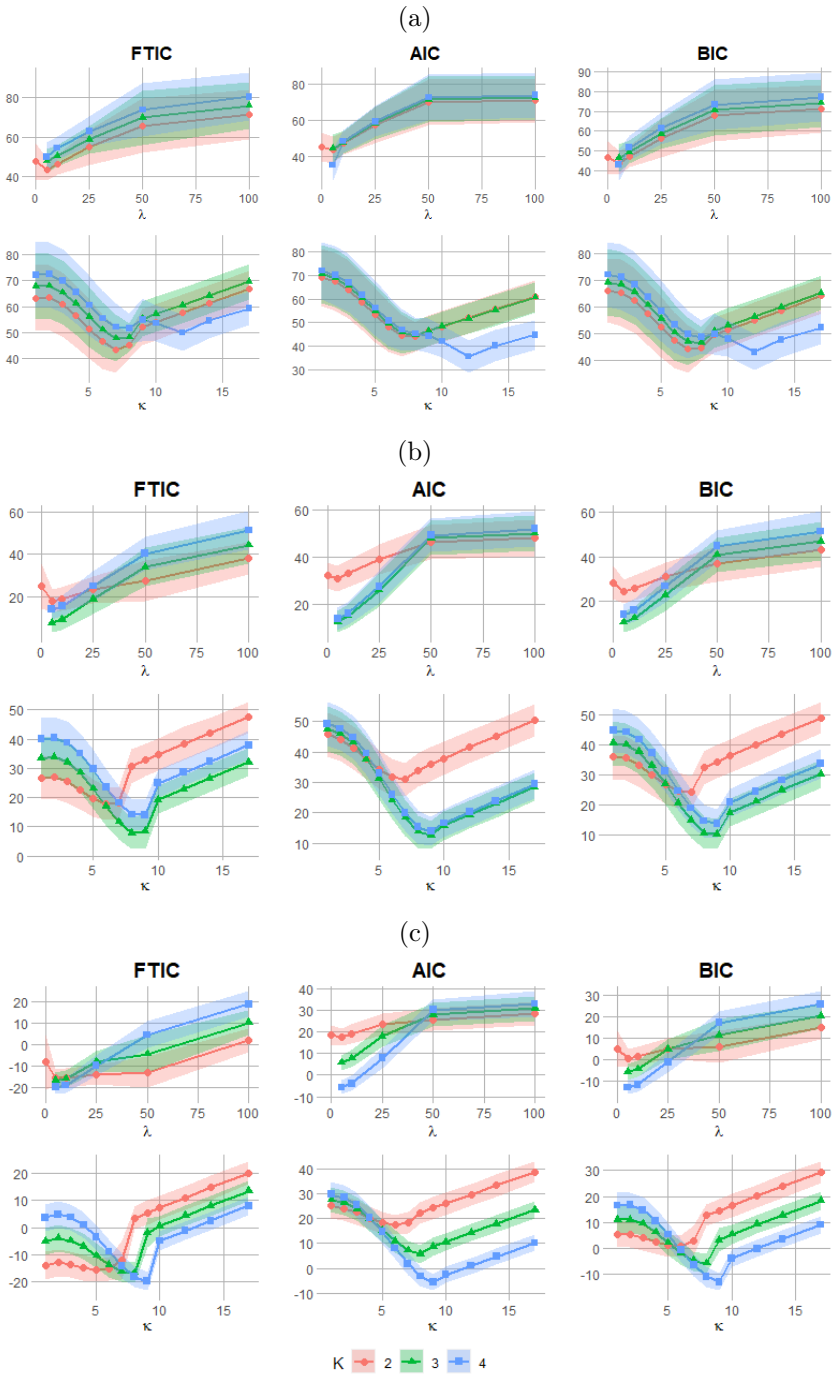


Figure 4.2: FTIC, AIC, and BIC of the SJM from 100 simulations each of length $T = 600$ for different values of λ and κ. Panel (a)–(c) depict the average ICs and their ± 2 standard deviation confidence bands when the true number of latent states (K_{true}) is equal to 2, 3, and 4, respectively.

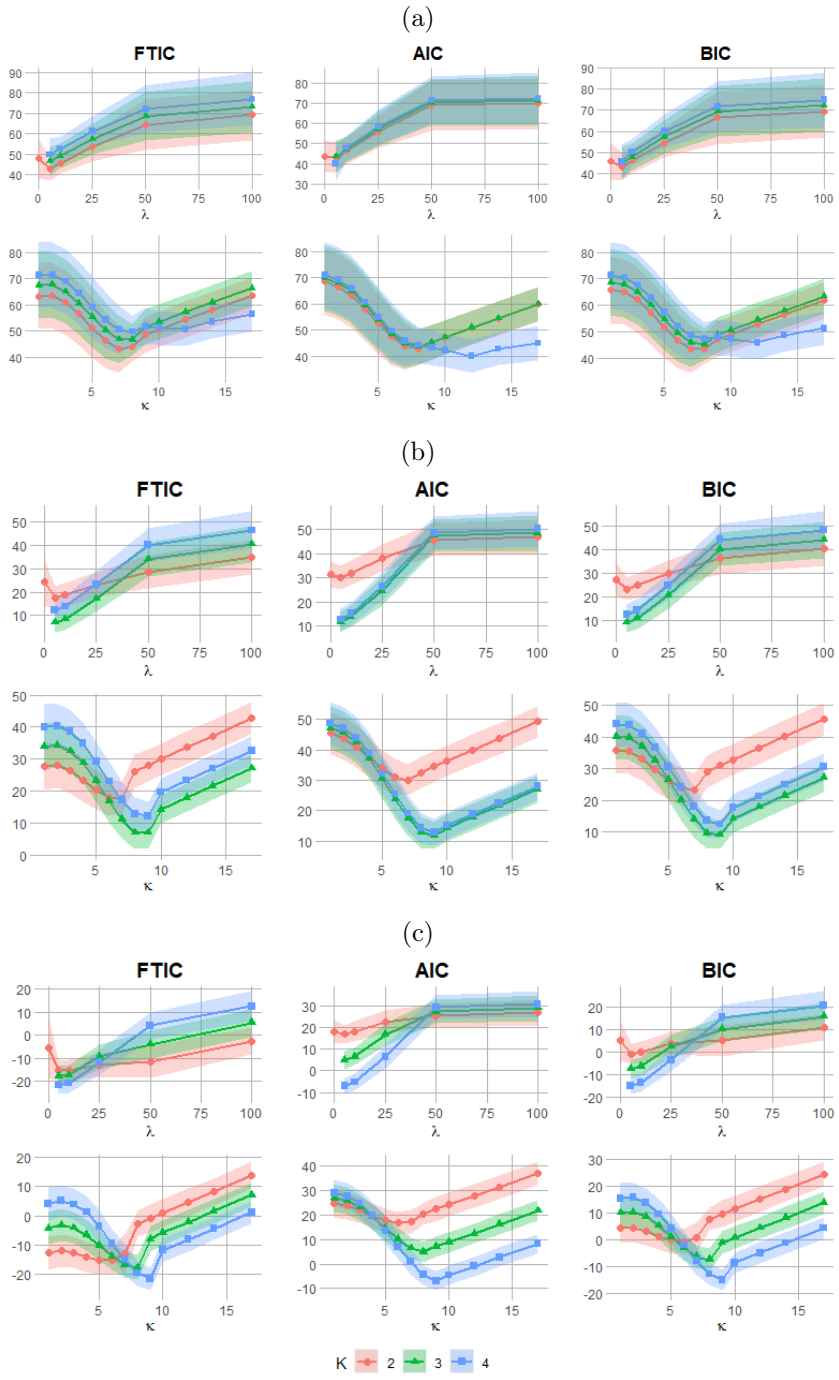


Figure 4.3: FTIC, AIC, and BIC of the SJM from 100 simulations each of length $T = 1,000$ for different values of λ and κ. Panel (a)–(c) depict the average ICs and their ± 2 standard deviation confidence bands when the true number of latent states (K_{true}) is equal to 2, 3, and 4, respectively.

Table 4.2: Simulation results for the minimum FTIC, AIC, and BIC and the corresponding λ, κ, and K for varying number of true latent states K_{true}, in the less persistent (a) and more persistent (b) HMM setups. Value refers to the averages across 100 simulations of the estimated value of the reported IC, and ARI($\{\hat{s}_t\}$) and ARI($\{\omega_i\}$) are the average ARIs computed between true and estimated sequences of states, and between true and estimated sequences of active features, respectively. TP (true positives) and FP (false positives) are the average numbers of correctly selected and wrongly selected features, respectively.

(a) Less persistent HMM

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_i\}$)	TP	FP
2	FTIC	42.98	5.00	7.00	2	1.00	0.55	63.45	0.00
	AIC	37.56	5.00	9.00	4	0.52	0.17	82.78	73.94
	BIC	42.41	5.00	9.00	4	0.52	0.17	82.78	73.94
3	FTIC	3.12	5.00	9.00	3	1.00	0.79	97.82	23.13
	AIC	10.72	5.00	9.00	3	1.00	0.79	97.82	23.13
	BIC	6.59	5.00	9.00	3	1.00	0.79	97.82	23.13
4	FTIC	-33.00	5.00	6.00	2	0.35	0.48	56.68	0.00
	AIC	-8.72	5.00	9.00	4	1.00	0.93	94.46	0.00
	BIC	-20.63	5.00	9.00	4	1.00	0.93	94.46	0.00

(b) More persistent HMM

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_i\}$)	TP	FP
2	FTIC	43.59	5.00	7.00	2	1.00	0.55	63.43	0.00
	AIC	43.52	5.00	8.00	2	1.00	0.59	91.83	54.96
	BIC	43.89	5.00	7.00	2	1.00	0.55	63.43	0.00
3	FTIC	11.16	5.00	9.00	3	1.00	0.78	98.02	27.99
	AIC	13.14	5.00	9.00	3	1.00	0.78	98.02	27.99
	BIC	12.06	5.00	9.00	3	1.00	0.78	98.02	27.99
3	FTIC	-11.67	10.00	9.00	4	0.99	0.93	94.43	0.00
	AIC	-4.29	5.00	9.00	4	1.00	0.92	94.54	2.00
	BIC	-8.28	5.00	9.00	4	1.00	0.92	94.54	2.00

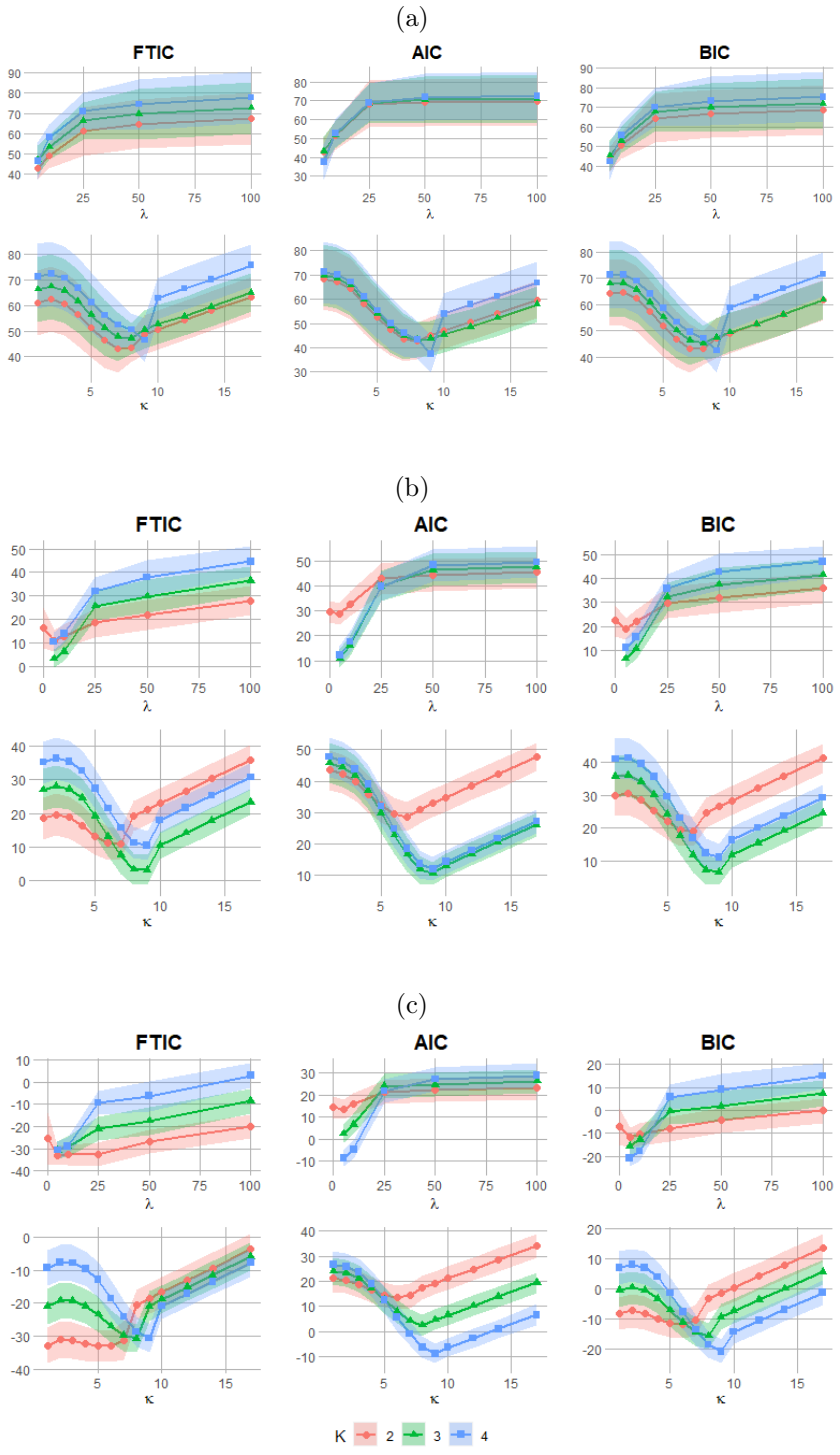


Figure 4.4: FTIC, AIC, and BIC of the SJM from 100 simulations each of length $T = 1,000$ of a less persistent HMM with transition probabilities given by Equations (4.6), (4.7) and (4.8). Panel (a)–(c) depict the average ICs and their ± 2 standard deviation confidence bands for different values of λ and κ when the true number of latent states (K_{true}) is equal to 2, 3, and 4, respectively.

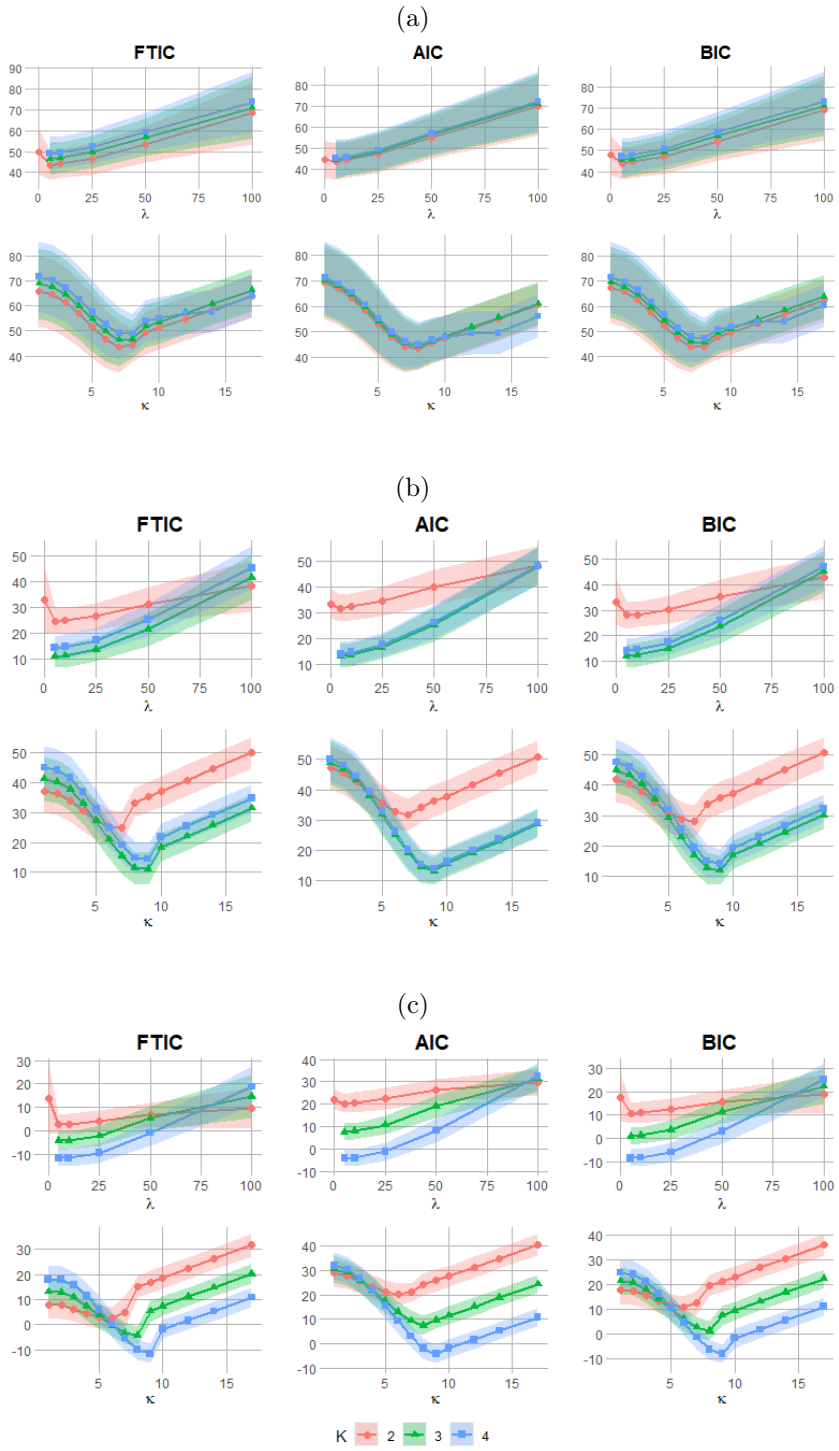


Figure 4.5: FTIC, AIC, and BIC of the SJM from 100 simulations each of length $T = 1,000$ of a more persistent HMM with transition probabilities given by Equations (4.9), (4.10) and (4.11). Panel (a)–(c) depict the average ICs and their ± 2 standard deviation confidence bands for different values of λ and κ when the true number of latent states (K_{true}) is equal to 2, 3, and 4, respectively.

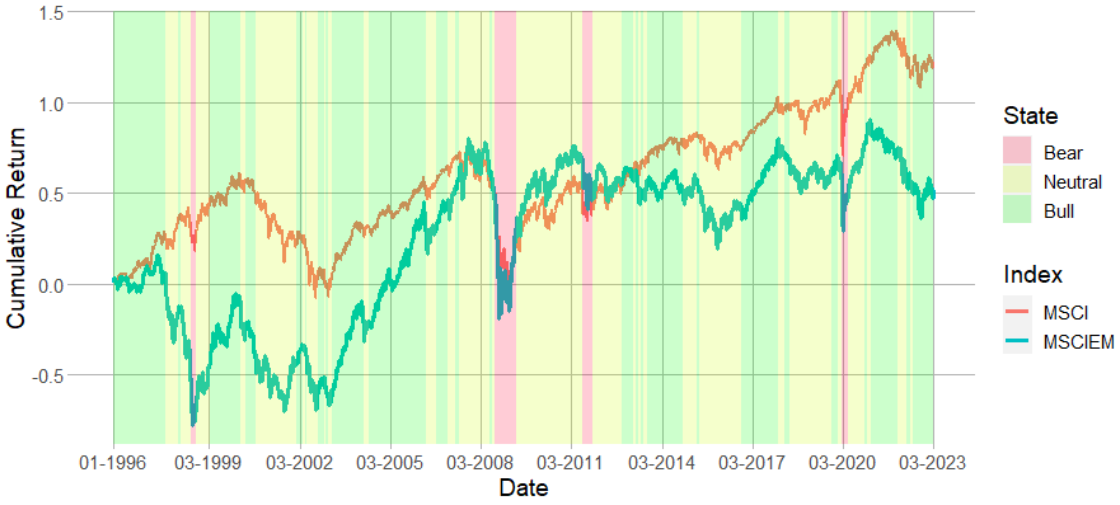


Figure 4.6: Cumulative log-returns of MSCI and MSCIEM with the state sequence of the best SJM as determined by FTIC.

Table 4.3: Selected features along with relative weights and state-conditional values. All values are expressed as percentage.

Feature	Weight	Bear	Neutral	Bull
$\mathrm{EMA}_{14}(\sigma_{\mathrm{MSCI}})$	4.10	2.36	0.94	0.64
$\mathrm{EMA}_{30}(\sigma_{\mathrm{MSCIEM}})$	4.07	2.42	1.17	0.82
$\mathrm{EMA}_{14}(\sigma_{\mathrm{MSCIEM}})$	4.00	2.54	1.15	0.78
$\mathrm{EMA}_{30}(\sigma_{\mathrm{MSCI}})$	3.93	2.21	0.97	0.67
$\rho_{30}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	3.91	-51.36	-42.20	-21.63
$g_{90}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	3.91	-20.07	-17.09	-10.12
$\rho_{90}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	3.80	-48.43	-40.84	-26.21
$\mathrm{EMA}_{7}(\sigma_{\mathrm{MSCI}})$	3.78	2.38	0.93	0.62
$\mathrm{EMA}_{7}(\sigma_{\mathrm{MSCIEM}})$	3.55	2.55	1.13	0.77
$g_{30}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	3.16	-23.32	-18.59	-9.54
$\mathrm{EMA}_{90}(\sigma_{\mathrm{MSCIEM}})$	3.07	2.04	1.19	0.90
$\rho_{14}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	3.05	-50.85	-42.19	-19.19
$\rho_{30}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	3.03	76.39	67.95	51.21
$g_{90}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	3.02	36.33	30.81	24.36
$\mathrm{EMA}_{2}(\sigma_{\mathrm{MSCI}})$	2.84	2.31	0.89	0.60
$g_{30}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	2.69	40.57	33.38	24.15
$\rho_{14}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	2.66	76.64	68.01	49.33
$\mathrm{EMA}_{90}(\sigma_{\mathrm{MSCI}})$	2.58	1.79	0.99	0.75
$\rho_{90}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	2.58	75.62	67.23	54.95
$\mathrm{EMA}_{2}(\sigma_{\mathrm{MSCIEM}})$	2.53	2.47	1.08	0.74
$\mathrm{EMA}_{90}(r_{\mathrm{MSCI}})$	2.31	-0.13	0.01	0.05
$g_{14}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	2.30	-26.40	-20.20	-9.14
$\mathrm{EMA}_{1}(\sigma_{\mathrm{MSCI}})$	2.26	2.24	0.86	0.58
$\rho_{7}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	2.04	-49.81	-41.66	-17.61
$\rho_{7}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	2.03	76.57	67.84	47.78
$\mathrm{EMA}_{90}(r_{\mathrm{MSCIEM}})$	1.94	-0.19	-0.00	0.05
$\mathrm{EMA}_{1}(\sigma_{\mathrm{MSCIEM}})$	1.93	2.39	1.05	0.71
$g_{14}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	1.68	40.65	35.38	24.99
$\rho_{90}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	1.49	-76.35	-73.77	-65.96
$\mathrm{EMA}_{30}(r_{\mathrm{MSCI}})$	1.48	-0.20	0.01	0.05
$g_{7}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	1.45	-25.33	-21.65	-7.96
$\mathrm{EMA}_{90}(r_{\mathrm{VIX}})$	1.44	0.36	0.02	-0.05
$\rho_{30}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	1.44	-79.61	-75.10	-66.04
$\rho_{14}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	1.06	-79.42	-75.22	-65.10
$g_{7}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	1.06	41.47	36.77	24.98
$\mathrm{EMA}_{30}(r_{\mathrm{MSCIEM}})$	1.05	-0.27	-0.01	0.06
$g_{90}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.86	-39.06	-33.80	-29.74
$\rho_{2}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	0.77	75.99	66.33	44.61
$g_{30}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.67	-41.45	-36.48	-31.93
$\rho_{7}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.62	-78.64	-74.73	-64.12
$g_{14}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.57	-43.75	-37.76	-31.92
$\rho_{2}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	0.55	-48.64	-39.14	-15.78
$\mathrm{EMA}_{14}(r_{\mathrm{MSCI}})$	0.54	-0.20	0.01	0.05
$\mathrm{EMA}_{30}(r_{\mathrm{VIX}})$	0.38	0.54	0.02	-0.06
$\mathrm{EMA}_{14}(r_{\mathrm{MSCIEM}})$	0.30	-0.25	-0.01	0.06
$\rho_{1}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	0.29	74.08	64.20	42.17
$g_{2}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	0.26	-24.86	-22.00	-7.65
$g_{7}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.23	-43.53	-39.44	-32.58
$\mathrm{RF}(r_{\mathrm{MSCIEM}})$	0.20	-0.47	-0.04	0.06
$\mathrm{RF}(r_{\mathrm{MSCI}})$	0.17	-0.45	-0.00	0.05
$g_{2}(r_{\mathrm{MSCI}}, r_{\mathrm{MSCIEM}})$	0.15	40.66	37.26	24.87
$\rho_{1}(r_{\mathrm{MSCIEM}}, r_{\mathrm{VIX}})$	0.14	-47.59	-36.87	-14.54
$\rho_{90}(v_{\mathrm{MSCIEM}}, r_{\mathrm{MSCIEM}})$	0.05	3.84	0.09	2.97
$\mathrm{EMA}_{7}(r_{\mathrm{MSCI}})$	0.02	-0.18	0.00	0.05
$\rho_{2}(r_{\mathrm{MSCI}}, r_{\mathrm{VIX}})$	0.01	-76.93	-72.11	-61.73

Table 4.4: Weights distribution and average state-conditional values by group of features. All values are expressed as percentage.

Group	Weight	Bear	Neutral	Bull
Volatility	38.64	2.31	1.03	0.72
Correlation with VIX	31.52	-48.77	-43.04	-30.42
Cross-correlation	19.96	59.54	52.29	37.58
Momentum	8.01	-0.26	0.00	0.05

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5:31–56.

Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4:313–337.

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

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *Test*, 23:433–465.

Bemporad, A., Breschi, V., Piga, D., and Boyd, S. P. (2018). Fitting jump models. *Automatica*, 96:11–21.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771.

Chen, Y., Mabu, S., Shimada, K., and Hirasawa, K. (2009). A genetic network programming with learning approach for enhanced stock trading model. *Expert Systems with Applications*, 36:12537–12546.

Das, S. and Biswas, A. (2023). Chapter four - Analyzing correlation between quality and accuracy of graph clustering. In Patgiri, R., Deka, G. C., and Biswas, A., editors, *Principles of Big Graph: In-depth Insight*, volume 128 of *Advances in Computers*, pages 135–163. Elsevier.

Dua, P. and Tuteja, D. (2021). Regime shifts in the behaviour of international currency and equity markets: A Markov-switching analysis. *Journal of Quantitative Economics*, 19:309–336.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:531–552.

Gerber, S., Markowitz, H. M., Ernst, P. A., Miao, Y., Javid, B., and Sargen, P. (2022). The Gerber statistic: A robust co-movement measure for portfolio optimization. *The Journal of Portfolio Management*, 48:87–102.

Guo, F., Chen, C. R., and Huang, Y. S. (2011). Markets contagion during financial crisis: A regime-switching approach. *International Review of Economics & Finance*, 20:95–109.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41:190–195.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–A review and recommendations for the practicing statistician. *Biometrical journal*, 60:431–449.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Springer Nature.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York, NY.

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

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). Time-series momentum. *Journal of Financial Economics*, 104:228–250.

Nystrup, P., Boyd, S., Lindström, E., and Madsen, H. (2019). Multi-period portfolio selection with drawdown control. *Annals of Operations Research*, 282:245–271.

Nystrup, P., Hansen, B. W., Madsen, H., and Lindström, E. (2015). Regime-based versus static asset allocation: Letting the data speak. *Journal of Portfolio Management*, 42:103–109.

Nystrup, P., Kolm, P. N., and Lindström, E. (2020a). Greedy online classification of persistent market states using realized intraday volatility features. *The Journal of Financial Data Science*, 2:25–39.

Nystrup, P., Kolm, P. N., and Lindström, E. (2021). Feature selection in jump models. *Expert Systems with Applications*, 184:115558.

Nystrup, P., Lindström, E., and Madsen, H. (2020b). Hyperparameter optimization for portfolio selection. *The Journal of Financial Data Science*, 2:40–54.

Nystrup, P., Lindström, E., and Madsen, H. (2020c). Learning hidden Markov models with persistent states by penalizing jumps. *Expert Systems with Applications*, 150:113307.

Nystrup, P., Madsen, H., and Lindström, E. (2017). Long memory of financial time-series and hidden markov models with time-varying parameters. *Journal of Forecasting*, 36:989–1002.

Rydén, T. (2008). EM versus Markov Chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3:659–688.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics*, 13:217–244.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–242.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:111–133.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:44–47.

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science*, 153:12–18.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

Yao, Y., Cao, Y., Zhai, J., Liu, J., Xiang, M., and Wang, L. (2020). Latent state recognition by an enhanced hidden Markov model. *Expert Systems with Applications*, 161:113722.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R.* CRC press, Boca Raton, FL.