Financial Innovation

RESEARCH

Open Access

Weighted-indexed semi-Markov model: calibration and application to financial modeling

Riccardo De Blasis*

*Correspondence:
r.deblasis@univpm.it

Department of Management, Marche Polytechnic University, Ancona, Italy

Abstract

We address the calibration issues of the weighted-indexed semi-Markov chain (WISMC) model applied to high-frequency financial data. Specifically, we propose to automate the discretization of the price returns and the volatility index by using four different approaches, two based on statistical quantities, namely, the quantile and sigma discretization, and two derived by the application of two popular machine learning algorithms, namely the k-means and Gaussian mixture model (GMM). Moreover, by comparing the Bayesian information criterion (BIC) scores, the GMM approach allows for the selection of the number of states of returns and index. An application to Bitcoin prices at 1-min and 1-s intervals shows the validity and usefulness of the proposed discretization approaches. In particular, GMM discretization is well suited for high-frequency returns, whereas the quantile approach works better for low-frequency intervals. Finally, by comparing the results of the Monte Carlo simulation, we show that the WISMC model, applied with the proposed discretization, can reproduce the long-range serial correlation of the squared returns, which is typical of the financial markets and, in particular, the cryptocurrency market.

Keywords: Semi-Markov, WISMC, Bitcoin, EWMA, k-means, GMM

JEL Classification: C63, C38, G17

Introduction

The general approach to studying financial time series is mostly based on applying econometric tools in time series analysis, in which the observed price is considered a noisy representation of an unobserved price. This approach is generally referred to as the macro-to-micro approach. However, in recent years, a new strand of literature has emerged. This new area deals with these problems by looking at the opposite perspective called the micro-to-macro approach, which directly models observable quantities and exploits point processes (Fodra and Pham 2015).

Among this new area of the literature, one of the first attempts to model financial time series using a semi-Markov chain is from D'Amico and Petroni (2012a), followed by an extension of the model by introducing a memory index (D'Amico and Petroni 2011).

Other authors have employed the semi-Markov process to model the limit order book (Swishchuk et al. 2017).

However, the approach that reported the best results is the weighted-indexed semi-Markov chain model (WISMC) by D'Amico and Petroni (2012b) and its multivariate extensions (D'Amico and Petroni 2018, 2021). The model has proven to reproduce important stylized facts of financial time series, such as first-passage-time distributions and the persistence of volatility. Moreover, it has also been employed in other applications. Specifically, D'Amico et al. (2018) applied the WISMC approach to model financial volumes, whereas D'Amico et al. (2020b) employed the model to study some risk measures in a high-frequency financial setting. In other fields, a simple indexed version of the model has been applied to analyze wind-power generation (D'Amico et al. 2020a).

The WISMC model can be regarded as a generalization of the semi-Markov chain model. Although the latter employs two random variables, namely, the observed price returns and the time between each price change, the WISMC includes a third variable that considers the history of the price returns and their intercurrent time, thus allowing for better reproduction of the observed quantities. However, in their original paper, D'Amico and Petroni (2012b) highlighted that applying the WISMC model to financial time series requires calibration of several parameters involved in the model. Mainly, we have to deal with converting continuous returns into a discrete state space. Moreover, the inclusion of an index that captures the history of the process requires further discretization. In D'Amico and Petroni (2012b), both conversions were based on visual inspection of the distribution of both processes, thus imposing a subjective choice. D'Amico et al. (2019) addressed the partition of the state space of an indexed Markov chain employing a change point approach.

In this study, we explore the possibility of automating the discretization of both price and index processes by testing the effectiveness of two simple discretizations, one based on quantiles and the other based on the returns standard deviation, and two algorithms taken from the machine learning literature, namely, the k-means and Gaussian mixture model (GMM). We included two machine learning algorithms because clustering and feature selection are two important research areas in applied financial research, especially given the complex distribution of financial data, and their respective literature is rapidly expanding. For example, Li et al. (2021) proposed an integrated cluster detection approach for financial applications, such as credit evaluation and fraud detection. Furthermore, Kou et al. (2021) employed machine learning algorithms to predict the bankruptcy of small and medium-sized enterprises (SMEs) using transactional data and payment network-based variables. Moreover, with automatic discretization, we can limit the discretion to the choice of the number of states. However, at the end of the paper, we show that using the GMM approach allows us to find the optimal number of states for both the returns and the index based on the Bayesian information criterion (BIC).

In addition, considering that the WISMC model has only been tested on stock markets, we apply the model to the cryptocurrency market in this study, particularly to the most recent Bitcoin prices from the Binance market, which is one of the most active cryptocurrency markets. The aim is to capture the typical stylized facts of this type of financial market, specifically the extremely high volatility inherent in Bitcoin prices, its high persistence, heavy tail behavior, and vulnerability to speculative bubbles. For

example, Hafner (2020) found evidence of bubbles and extreme volatility by testing 11 of the largest cryptocurrencies. Meanwhile, Bariviera et al. (2017) analyzed the stylized facts of the Bitcoin market and found long memory in returns time series, indicating price predictability and market inefficiency. Moreover, Tan et al. (2020) assessed the volatility of 102 cryptocurrencies using Garman-Klass volatility measures, demonstrating the complexity of understanding such assets.

The application of the WISMC model to the Bitcoin market shows that the algorithms are useful in the discretization of both the returns process and index. More specifically, the quantile approach works better for lower-frequency data, whereas the GMM approach is better suited for higher-frequency returns. In addition, the BIC score of the GMM approach allows for the automation of choosing the number of states.

The remainder of this paper is organized as follows. "The model" section describes the WISMC model's theory, whereas "Discretization algorithms" section introduces four discretization approaches. "Application to financial data" section explores the challenges of the calibration process and shows the data along with the discretization results. Finally, "Conclusion" section concludes the paper.

The model

First, we introduce the semi-Markov processes from which the weighted-indexed semi-Markov process is derived. They were first proposed by Levy (1954) and Smith (1955) independently and further studied by Pyke (1961a, 1961b) and Çinlar (1975). Subsequently, they found applications in many fields, from industrial to financial markets, and the theory has been further implemented and expanded (see, e.g., Vasileiou and Vassiliou 2006; Swishchuk et al. 2017; Pasricha et al. 2020). For an in-depth analysis, we refer the readers to Janssen and Manca (2006) and Barbu and Limnios (2009).

Semi-Markov processes can be viewed as a generalization of renewal processes and the Markov chain. Let us consider a finite state space $E = \{1, ..., k\}$ and a probability space $(\Omega, \mathcal{F}, P)$. The two random variables

$$X_n : \Omega \to E \qquad T_n : \Omega \to \mathbb{R}_+,$$

where $n \in \mathbb{N}$ and $0 = T_0 < T_1 < T_2 < \ldots$ form a *Markov renewal process* (X, T) with a state space $E \times R_+$ if

$$\begin{aligned} &\mathbb{P}(X_{n+1} = j, T_{n+1} - T_n \leq t | X_0, \ldots, X_n; T_0, \ldots, T_n) \\ &= \mathbb{P}(X_{n+1} = j, T_{n+1} - T_n \leq t | X_n) \quad a.s., \quad \forall n \in \mathbb{N}, j \in E, t \in \mathbb{R}_+. \end{aligned} \tag{1}$$

Assuming that the process is temporally homogeneous, the probability

$$\mathbb{P}(X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i) = Q(i, j, t) \tag{2}$$

is independent of n, and Q is called a *semi-Markov kernel*. In general, $Q(i, j, 0) = 0$, $\forall i, j \in E$.

For each pair (i, j),

$$\lim_{t \to +\infty} Q(i, j, t) = P(i, j), \tag{3}$$

where $P(i,j) \geq 0$ and $\sum_{j \in E} P(i,j) = 1, i,j \in E$. The quantities $P(i, j)$ are the transition probabilities of the Markov chain, $\{X_n\}_{n \in N}$, with state space E.

Moreover, we can define the conditional waiting time distribution function as

$$G(i,j,t) = \mathbb{P}(T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j), \tag{4}$$

which can be computed as

$$G(i,j,t) = \frac{Q(i,j,t)}{P(i,j)}, \tag{5}$$

with the convention that $G(i,j,t) = 1$ if $P(i,j) = 0$, and it can be proven that the increments $T_{n+1} - T_n$ are conditionally independent given the Markov chain X_n (see, e.g., Çinlar 1975).

In particular, when the state space E is composed of a single point, the increments are independent and identically distributed nonnegative random variables, and we obtain a *renewal process*.

We can now define the *semi-Markov process* with state space E and transition kernel $Q(i, j, t)$ as a continuous-time parameter process:

$$Y_t = X_n \text{ for } t \in [T_n, T_{n+1}). \tag{6}$$

This process can be considered the state at time t of a system that moves from one state to another with random sojourn times in between (Çinlar 1975). The sojourn interval $[T_n, T_{n+1})$ represents a random variable with a distribution that depends on the state being visited X_n and the next state to be visited X_{n+1}.

The semi-Markov process is called so because it cannot be fully considered a Markovian process as it is not a memoryless process. In contrast, it follows the Markov property only at jump instants. In addition, when sojourn times are exponentially distributed, the semi-Markov process becomes a continuous-time Markov chain. Instead, we obtain a discrete-time Markov chain if we ignore time variables.

The semi-Markov process can be further extended by including the memory of the process using high-order semi-Markov processes (see, e.g., (Limnios and Oprian 2003; D'Amico et al. 2013). However, this method requires the estimation of several parameters. A more parsimonious model considers the dependence of the semi-Markov process on a third variable that considers the history of the process. This approach was initially considered in D'Amico (2011) and was further extended to financial applications in D'Amico and Petroni (2011, 2012b).

Let U_n be a stochastic process with the values in $\mathbb{R}$. This random variable represents the indexing process that stores the historical information of the semi-Markov process and can be expressed as D'Amico and Petroni (2021)

$$U_n(\boldsymbol{\theta}) = \sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} f(X_{n-1-k}, T_n, a, \boldsymbol{\theta})) + f(X_n, T_n, T_n, \boldsymbol{\theta}), \tag{7}$$

where $f : E \times \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ is a Borel measurable bounded function and U_0 is known and non-random. The size of the vector of the parameters $\boldsymbol{\theta}$ depends on the chosen function f.

Process Y_t is said to be a *WISMC* if, $\forall n \in \mathbb{N}$, the following equality holds true:

$$\begin{aligned} &\mathbb{P}(X_{n+1} = j, T_{n+1} - T_n \leq t | X_0, \ldots, X_n; T_0, \ldots, T_n; U_0, \ldots, U_n) \\ &= \mathbb{P}(X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i, U_n = v) := Q(i, j, t, v), \end{aligned} \tag{8}$$

where the function Q is called the *indexed semi-Markov kernel*.

Condition (8) states that to assess the probability of the next state of the process, we only need knowledge of the last state i and the last value of the indexing process U_n. Therefore, the triple process $\{X_n, T_n, U_n\}$ describes the system corresponding to any jump time T_n. Note that if the indexed semi-Markov kernel is constant in v, then it degenerates into a semi-Markov kernel, and the WISMC process becomes a semi-Markov process.

Moreover, for each pair (i, j) and each value of the index, we have $Q(i, j, 0, v) = 0$ and

$$P(i, j, v) = \mathbb{P}(X_{n+1} = j | X_n = i, U_n = v). \tag{9}$$

The quantities $P(i, j, v)$ are the transition probabilities of the Markov chain, $\{X_n\}_{n \in N}$, with state space E. These differ from the probabilities in (3) because they depend on the index level.

Moreover, the conditional waiting time distribution function includes dependence on the index level:

$$G(i, j, t, v) = \mathbb{P}(T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j, U_n = v). \tag{10}$$

Discretization algorithms

We encounter several calibration issues when applying the WISMC or semi-Markov model to financial data. The first step at the beginning of the application is the discretization of the price return, as the WISMC model is defined in discrete state space. In contrast, the returns we observe in real life are continuous. In their application, D'Amico and Petroni (2012b) relied on arbitrary discretization based on the visual observation of the returns histogram. Unfortunately, this approach cannot be used for automated routines. Therefore, we introduce four algorithms to deal with this discretization of price returns. The first two approaches are simple, as they are based on the statistical properties of returns. The first merely consists of splitting the observations into k quantiles, where k is the number of states. We refer to this approach as *quantile* discretization. Subsequently, by selecting the splitting point, we built the edges of the states. Although this discretization is easy to implement, it may present some issues. For example, if we select a high number of quantiles when dealing with a highly leptokurtic distribution, which is typical of a financial series, observations with a high frequency, typically the zero return, might be distributed between two contiguous states, thus resulting in non-unique state edges.[1]

The second approach was proposed by De Blasis (2020) for the return series, and we refer to it as *sigma* discretization. The idea was to select the width of the states as the standard deviation of the observations. Then, based on the number of states and

[1] An example is given in "Application to financial data" section.

Table 1 Examples of the sigma approach with odd and even numbers of states

State 1	State 2	State 3	State 4
$[min, -\sigma)$	$[-\sigma, 0)$	$[0, \sigma)$	$[\sigma, max]$

State 1	State 2	State 3	State 4	State 5
$[min, -2\frac{\sigma}{2})$	$[-2\frac{\sigma}{2}, -\frac{\sigma}{2})$	$[-\frac{\sigma}{2}, \frac{\sigma}{2})$	$[\frac{\sigma}{2}, 2\frac{\sigma}{2})$	$[2\frac{\sigma}{2}, max]$

centering them to zero, that is, the null return, we build the edges of the states. If the number of states is odd, then the central bin contains all zero returns together with smaller returns within a half standard deviation radius from the zero return. Then, departing from this central state, the other bins are defined as the one standard deviation distance from the others, leaving the extreme states up to the returns' minimum and maximum values. In the case of an even number of states, the central zero return state is omitted, and we have only positive and negative return states.[2] Table 1 shows the concept of both odd and even numbers of states. This approach is well designed to reproduce symmetric distributions of continuous returns, especially when choosing an odd number of states, as it can provide an immediate idea of the direction of returns and includes a portion of the market noise within the central bin.

The other two discretization approaches employ two popular clustering algorithms: k-means and GMM. The *k-means* algorithm is a simple unsupervised algorithm developed independently by Sebestyen (1962) and MacQueen (1967). The idea is to partition the observations so that the within-cluster sum of squares is minimized using an iterative algorithm.[3] Once we define the number of clusters k, that is, the states of the WISMC model in our specific application, the algorithm returns the discretization with the association of each continuous return to a specific state, thereby minimizing the variance within the clusters. The advantage of this approach is that it is completely endogenous and follows an empirical distribution of price returns. By contrast, with many observations, the k-means algorithm can result in slow convergence. To speed up the algorithm, we use a variation called the mini-batch k-means introduced by Sculley (2010), which lowers the computational cost by using random samples of the full dataset, thus reducing the number of distances to compute at the cost of a lower quality of the clusters.

Because the *k-means* algorithm presents some limitations, see, for example, Li et al. (2021), we include a fourth discretization performed using the GMM algorithm, which is based on the assumption that the observations are generated by a mixture of Gaussian distributions with unknown parameters. The first studies in this direction were proposed by Wolfe (1963) and Scott and Symons (1971) and further studied by many other authors (see, e.g., Banfield and Raftery 1993; Fraley and Raftery 2002).[4] Let us assume that the observations $\{z_1, ..., z_t\}$ (i.e., the price returns) are realizations of a random vector $Z \in \mathbb{R}$

[2] The zero returns can be included in either the positive or negative state.

[3] For a review of the k-means clustering methods we refer the reader to Steinley (2006).

[4] For a comprehensive review of the finite mixture clustering, we refer the reader to Bouveyron and Brunet-Saumard (2014) and Bouveyron et al. (2019).

and the unobserved state labels $\{y_1, ..., y_t\}$ are realizations of a random variable $Y \in E$. If we denote g as the probabilistic density function of Y, then the GMM is

$$g(z;\theta) = \sum_{i=1}^{k} \pi_i \phi(z;\theta_i) \tag{11}$$

where π_i is the mixture proportion with the constraint $\sum_{i=1}^{k} \pi_i = 1$ and $\phi(z;\theta_i)$ is the Gaussian density with parameter $\theta_i = (\mu_i, \sigma_i)$, which are generally estimated using the expectation-maximization (EM) algorithm, proposed by Dempster et al. (1977). One of the advantages of the GMM algorithm is that it allows us to select the optimal number of clusters based on the BIC criterion.

Application to financial data

The application to financial data requires the formalization of the functional form of the index $U_n(\boldsymbol{\theta})$. D'Amico and Petroni (2011) initially proposed using a moving average of the squared process, $(X_n)^2$. Taking the square of the returns, the authors introduced the dependence of the process dynamics on volatility, which is an observed stylized fact in financial markets. In a later study, the authors opted for an exponentially weighted moving average (EWMA) of the squares of returns (D'Amico and Petroni 2012b). Using EWMA changes the function to

$$f(X_{n-1-k}, T_n, a, \lambda) = \frac{\lambda^{T_n - a} X^2_{n-1-k}}{\sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \lambda^{T_n - a}}. \tag{12}$$

The output values of the EWMA function in (12) were continuous. Therefore, similar to price returns, the index values need to be discretized into finite states using the proposed discretization algorithms in "Discretization algorithms" section.

Finally, to test the validity of the proposed approach for discretization, we performed a Monte Carlo simulation. We simulated a WISMC process using the following algorithm (D'Amico and Petroni 2012b):

1. set $n = 0$, $X_0 = i$, $T_0 = 0$, $U_0 = v$, horizon time $= T$
2. given X_n and U_n, sample X from $P(i, j, v)$ and set X_{n+1}
3. given X_n and X_{n+1}, sample W from $G(i, j, t, v)$ and set $T_{n+1} = T_n + W$
4. set U_{n+1} using (7) and (12)
5. if $T_{n+1} \geq T$ stop, else set $n = n + 1$ and go to 2.

To estimate the transition probability matrices $P(i, j, v)$ and conditional waiting time distribution $G(i, j, t, v)$, we refer the readers to Appendix B in D'Amico and Petroni (2018)

We then verify whether the simulated series follows the long-range serial correlation of the squared returns, which is typical of the financial returns series. We recall the autocorrelation function of the squared returns:

$$\Sigma(\tau) = \frac{Cov(Y^2(t+\tau), Y^2(t)}{Var(Y^2(t)}, \tag{13}$$

Table 2 Summary statistics of the Bitcoin price (USD) and log-returns (%) for 1-min and 1-s intervals

	1-min interval		1-s interval	
	Price	**Log-returns**	**Price**	**Log-returns**
Obs	524680		634172	
Mean	47501	0.0000	38269	0.0000
Std	9321	0.1160	1361	0.0153
Min	28868	− 5.9738	34330	− 0.3996
25%	39395	− 0.0519	37639	− 0.0016
50%	47320	− 0.0001	38435	0.0000
75%	56074	0.0507	39040	0.0008
Max	69000	4.8604	44219	0.4718
Skewness		− 0.0266		0.3774
Kurtosis		84.6502		30.4025

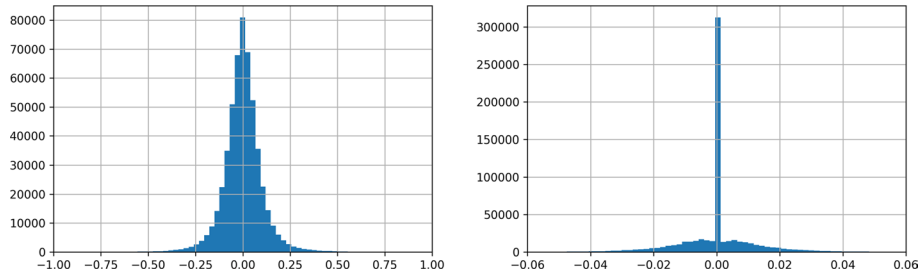

Fig. 1 On the left: histogram of the percentage 1-min returns series. Data from 1 March 2021 to 28 February 2022. The x-axis is limited between − 1% and 1%. On the right: histogram of the percentage 1-s returns series. Data from 21 February 2021 to 28 February 2022. The x-axis is limited between -0.06% and 0.06%

where Y is the process of returns and τ is the time lag. We estimate $\Sigma(\tau)$ for the real and simulated returns and compute the root mean square error (RMSE) and mean absolute error (MAE) to compare the use of different parameter estimations.

We tested the validity of the discretization algorithm on Bitcoin spot data sourced from the Binance public website.[5] We specifically selected Bitcoin data because the cryptocurrency market is open 24/7; thus, there are no gaps in the time series. In addition, we chose the Binance exchange because it is the largest cryptocurrency exchange in the world and is less subject to market manipulation (De Blasis and Webb 2022).

Following the approach of D'Amico and Petroni (2012b), we sample the price returns at 1-min intervals using Bitcoin data from March 1, 2021, to February 28, 2022. In addition, we test the application on 1-s interval returns with data ranging from February 21, 2022, to February 28, 2022. The date ranges vary because we aim to have a roughly similar number of observations in both samples. The summary statistics of the percentage log-returns are reported in Table 2. We observe a zero return on average with a standard deviation of 0.116% and 0.0153% for the 1-min and 1-s intervals, respectively. Both return distributions appear to be symmetric and follow the typical financial return

[5] Available at https://data.binance.vision Accessed March 1, 2022.

Table 3 Extreme values of the states bins for the four discretization approaches with 5 states applied to the Bitcoin returns

	State -2	State -1	State 0	State 1	State 2
Panel A: 1-min interval					
quantile	[− 5.974,− 0.066)	[− 0.066,− 0.018)	[− 0.018,0.017)	[0.017,0.065)	[0.065,4.860]
sigma	[− 5.974,− 0.174)	[− 0.174,− 0.058)	[− 0.058,0.058)	[0.058,0.174)	[0.174,4.860]
k-means	[− 5.974,− 0.130)	[− 0.130,− 0.022)	[− 0.022,0.066)	[0.066,0.247)	[0.247,4.860]
GMM	[− 5.974,− 0.065)	[− 0.065, 0.076)	[0.076,0.261)	[0.261,0.325)	[0.325,4.860]
Panel B: 1-s interval					
quantile	[− 0.4,− 0.005)	[− 0.005, 0.000)	[0.000,0.000)	[0.000,0.005)	[0.005,0.472]
sigma	[− 0.4,− 0.023)	[− 0.023,− 0.008)	[− 0.008,0.008)	[0.008,0.023)	[0.023,0.472]
k-means	[− 0.4,− 0.027)	[− 0.027,− 0.006)	[− 0.006,0.011)	[0.011,0.043)	[0.043,0.472]
GMM	[− 0.4,− 0.034)	[− 0.034,− 0.003)	[− 0.003,0.003)	[0.003,0.036)	[0.036,0.472]

distribution, with high excess kurtosis and fat tails, as shown in Fig. 1. In addition, the 1-s distribution presents a very high frequency around the null return.

As described in "Discretization algorithms" section, we discretize the continuous returns using four different approaches: quantile, sigma, k-means, and GMM discretization. The only discretion is left to the choice of the number of states, which, in our application, is set at three and five. The 4-state returns discretization is excluded from the analysis as an odd number of states would better follow the typical shape of the financial returns, which presents an almost symmetric distribution and a high frequency around the zero return. For space reasons, we report only the results of the 5-state discretization, which, for the sigma discretization, is identified by one central state representing the zero return surrounded by two positive and two negative states, corresponding to positive and negative returns, respectively. Table 3 lists the edges of each discretization bin for the four approaches. Panel A shows the discretization for the 1-min interval returns, whereas Panel B reports the edges of the bins for the 1-s interval returns. Note that the quantile discretization in this latter case fails because there is no way to attribute the continuous returns to State 0, State −1, or State 1. Therefore, we excluded this case from the subsequent analysis.

The results of the return discretization are also presented in Fig. 2, which shows the histograms built from the bins defined in Table 3. Quantile discretization is excluded from the charts as it results in a flat histogram. All discretizations present the highest frequency around the zero return; however, only the sigma discretization is symmetric around the zero return by construction. The k-means and GMM discretization of the 1-min returns appear to be asymmetric, whereas the distribution results are more balanced when using the 1-s returns. Moreover, the fourth state of the GMM discretization at the 1-min interval is minimal compared to the other states, which could result in a biased application. To this extent, we must highlight that the use of different discretizations leads to different distributions of WISMC processes, Y_t, which could be in different states simultaneously for different discretizations.

Once the returns are filtered into discrete states, we compute the index using the EWMA function. This stage requires calibration of the λ parameter using the technique discussed in D'Amico and Petroni (2012b) by minimizing the RMSE or MAE of the autocorrelation function of the simulated and real squared returns. However, as reported by

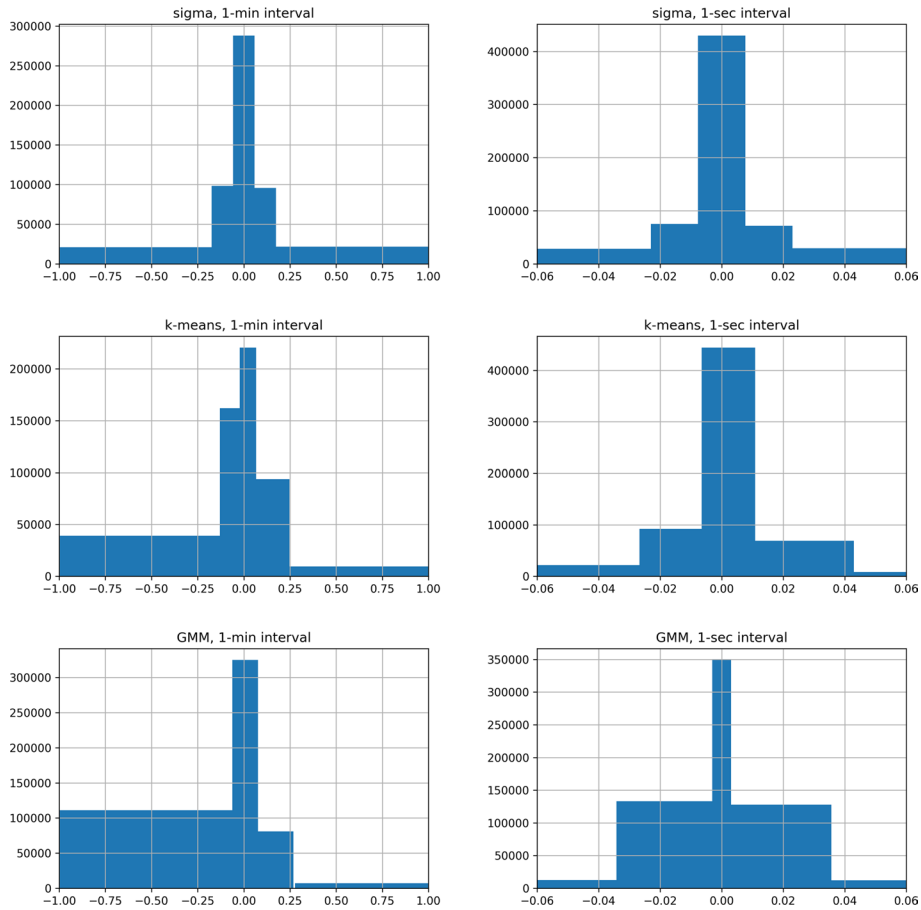

Fig. 2 Sigma (top), k-means (mid) and GMM (bottom) discretizations of the percentage 1-min (left) and 1-s (right) returns series

the authors and tested in our samples, the optimum is reached when λ varies between 0.95 and 0.99, and the overall RMSE or MAE values do not change visibly within that range. Therefore, following D'Amico and Petroni (2012b), we fixed $\lambda = 0.97$ for all our analyses, focusing our results mainly on the discretization algorithms. Furthermore, we note that when $\lambda = 1$, the EWMA function reduces to the moving average index proposed in D'Amico and Petroni (2011).

As stated earlier, the index has values in $\mathbb{R}$; therefore, it must be discretized like returns. D'Amico and Petroni (2012b) discretize the index into five states, specifically low, medium-low, medium, medium-high, and high volatility, choosing manual bins based on the visual observation of the distribution. By contrast, we employ the discussed discretization algorithms. We exclude only the sigma approach because the distribution of the index is not always symmetrical. Moreover, we did not limit the index discretization to five states.

Table 4 presents the RMSE values for comparing the simulated and real autocorrelation values of the WISMC process. The MAE values are not reported for space reasons; however, they are equivalent to the RMSE values. The table reports two combinations of returns/index discretization, that is, 3-state returns and 3-state index, and 5-state returns and 5-state index. The 3-state GMM discretization for 1-min interval returns is

Table 4 RMSE of the autocorrelation of the squared returns for different combinations of returns/index discretization and number of states of the chain and of the index. Returns discretization is indicated by rows, index discretization by columns

	1-min interval			1-s interval		
	Quantile	***k-means***	***GMM***	***Quantile***	***k-means***	***GMM***
Panel A: 3-state returns; 3-state index						
Quantile	**0.015**	0.0194	0.0189	0.0385	0.0673	0.052
Sigma	0.0187	0.0279	0.0291	0.044	0.0454	0.0464
k-means	0.028	0.0269	0.0309	0.0637	0.0591	0.0465
GMM	–	–	–	**0.01**	**0.0103**	**0.0103**
Panel B: 5-state returns; 5-state index						
Quantile	**0.0065**	0.0137	0.0185	–	–	–
Sigma	0.0554	0.0124	0.0312	0.0525	0.0488	0.0485
k-means	0.0332	0.0212	0.0271	0.0686	0.0461	0.0417
GMM	0.066	0.0271	0.0414	0.0725	**0.0374**	0.0468

Bold values indicate the combination of returns/index discretization with the lowest RMSE

not reported, as the algorithm resulted in a 2-state discretization. Similarly, the 5-state quantile discretization for the 1-s interval has not been reported because of the ambiguity of the state attribution, as described previously. The results show that quantile/quantile discretization is better suited for lower frequency intervals, whereas GMM/GMM or similarly GMM/k-means discretization works better at higher frequencies. Overall, the quantile/quantile with five states applied to the 1-min interval appears to be the best fit. In addition, we note that when using GMM discretization for the returns, the discretization of the index ceases to be relevant, leaving discretion over the choice of the algorithm. The results also show that the sigma and k-means discretization for the returns do not produce good results compared to the other two approaches.

In addition, to better understand the effect of the discretization approaches, we plotted the autocorrelation function of the simulated WISMC process using the best combinations from our results and compared it to the autocorrelation function of the observed data. Figure 3 clearly shows that the 5-state quantile/quantile discretization applied to the 1-min interval data performs much better than the 3-state quantile-quantile approach. However, we note a slight deviation between the simulated and real autocorrelation at low lags; more specifically, the simulated autocorrelation is underestimated up to the 20th lag. In contrast, the 3-state GMM/GMM discretization applied to 1-s interval data performs better than the 5-state GMM/k-means approach, which is the worst performer overall. In the GMM/GMM case, the simulated autocorrelations deviate from the real ones only for high lag values. Thus, this discretization better captures the short autocorrelation.

The presented results depend on the choice of the number of states for returns and index discretization. However, because one of the advantages of GMM discretization is the possibility of using the BIC score to choose the number of states and considering that the GMM works well for higher frequencies, we automate the selection of the number of states using the BIC score and apply this methodology only to the 1-s interval returns. First, we compute the BIC score for the return discretization and choose the optimal number of states;

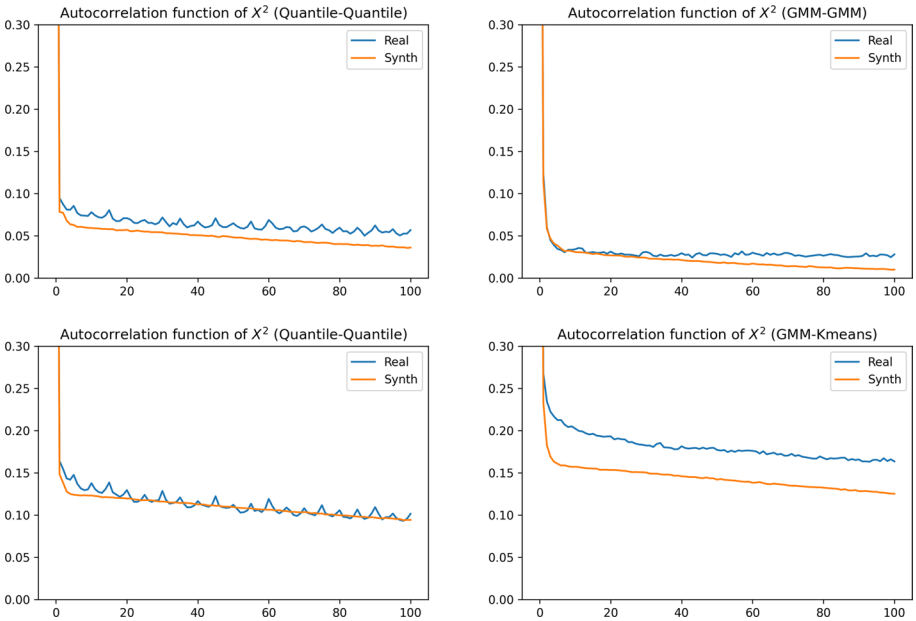

Fig. 3 Autocorrelation function of the simulated WISMC process against the real process. The discretization combinations are: quantile 3-state returns, quantile 3-state index 1-min interval (top-left); GMM 3-state returns, GMM 3-state index 1-sec interval (top-right); quantile 5-state returns, quantile 5-state index 1-min interval (bottom-left); GMM 5-state returns, k-means 5-state index 1-sec interval (bottom-right)

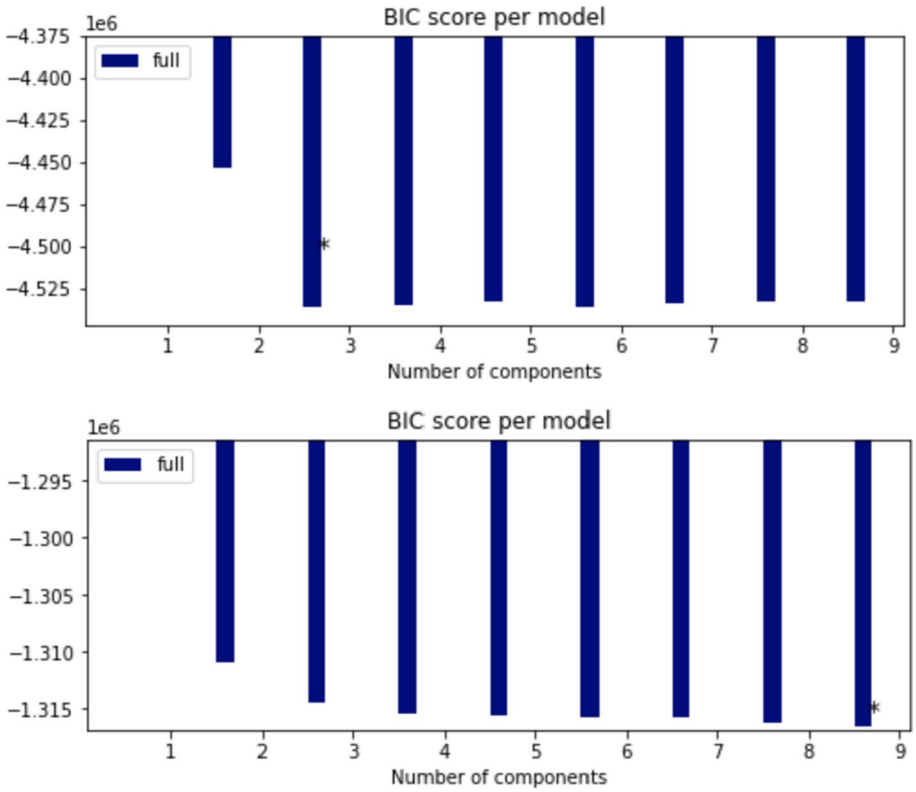

Fig. 4 States selection based on the BIC scores. Returns discretization on top and index discretization at the bottom

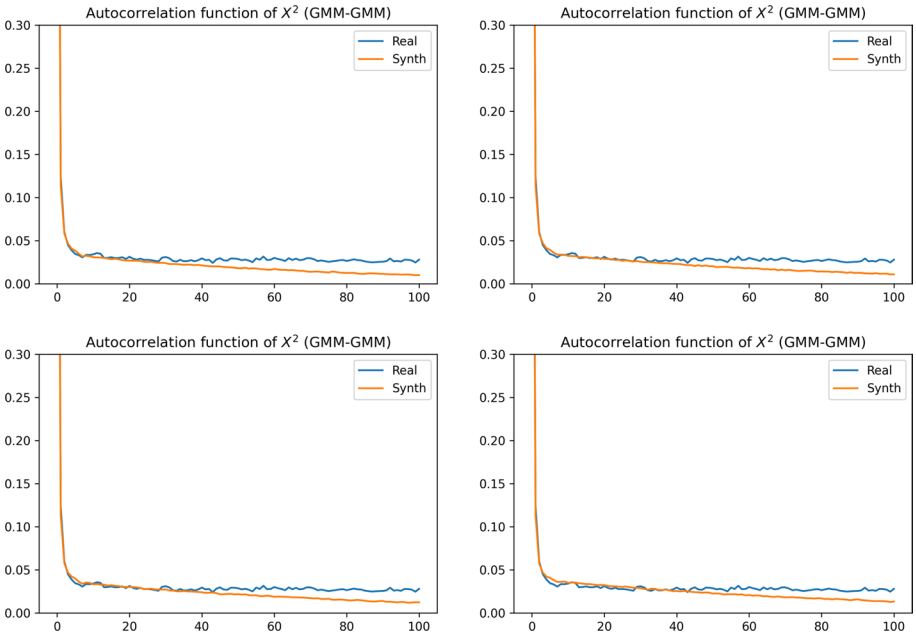

Fig. 5 Comparison of the GMM index discretization given a 3-state GMM Returns discretization. Returns at 1-s interval. 3-state index (top-left), 4-state index (top-right), 5-state index (bottom-left), 9-state index (bottom-right)

then, given the selected number of states for the returns, we compute the BIC score for the index discretization. The state selection is shown in Fig. 4, where the top chart indicates the optimal number of states for the returns, and the bottom chart indicates the optimal number of states for the index. The return discretization clearly reports the best score for the 3-state GMM approach, and this result appears to be in line with the RMSE results, where the 3-state discretization performed better than the 5-state one. Therefore, we fixed the number of states for the returns to three and proceeded with selecting the number of states for the index. In this case, we cannot directly choose the optimal BIC because its values appear to decline with the increment of the states. Note that adding states will result in estimating additional parameters, such as transition probabilities and sojourn time distribution. Therefore, we employ the *elbow method* to select the optimal score. We observed a significant drop in the BIC score from two to three states, followed by another smaller drop when four states were reached. Subsequently, from four to nine states, the decrease is reduced. Thus, we can easily select a 4-state GMM discretization for the index as a good trade-off between improving model performance and reducing the number of parameters to be estimated. Figure 5 compares the autocorrelation functions of both the simulated and real WISMC processes between the 3-, 4-, 5-, and 9-state index discretization. In all cases, short-run autocorrelation was well-fitted by the simulated data. However, we note that the 4-state discretization performs slightly better than the 3-state one, but adding more states to the index discretization does not significantly improve the performance of the model.

Conclusion

We proposed new calibration approaches to the WISMC model by D'Amico and Petroni (2012b). Specifically, we tested four different discretization methods for price returns: quantile, sigma, k-means, and GMM discretization. In addition, we use the same approaches, excluding sigma discretization, to discretize the volatility index, which represents the core part of the WISMC model. We tested different combinations of returns/index discretization on Bitcoin prices and found that the quantile/quantile approach works better for lower-frequency returns, whereas the GMM/GMM approach is better suited for higher-frequency data.

Moreover, we tested different combinations of number states for returns and indices. However, although selecting the number of states is generally left to the researcher's discretion, we showed that this choice could be automated when using GMM discretization. We propose selecting the number of states for the returns and the index based on the BIC score. The results reported by the comparison of the autocorrelation functions show that this methodology could be useful when implementing the WISMC model for high-frequency financial data. Overall, the model, with the inclusion of the automation of the discretization of the returns and volatility index, can reproduce the long-range serial correlation typical of financial markets.

This study presents some limitations that should be addressed in future research. For example, the model was validated by testing its ability to reproduce the autocorrelation of a financial time series. However, future studies could address other applied problems, such as price prediction, option pricing, and market and credit risk assessment, which are important problems in financial applications. Moreover, the choice of cryptocurrencies, which are financial assets traded 24/7 without breaks, helped reduce potential problems derived from trading halts due to possible price jumps. Therefore, a test of different assets is required to prove the validity of the model under different trading conditions. Further research might compare the model's results with other macro-to-micro approaches, such as the generalized autoregressive conditional heteroskedasticity model. Finally, the discretization approaches might be extended, employing more advanced clustering algorithms from the machine learning literature, see, for example, Li et al. (2021), and evaluate their performances with a multiple criteria decision making approach as in Kou et al. (2014).

Acknowledgements
The authors would like to thank the anonymous reviewers and participants at the FFEA2022 and SMTDA2022 conferences for their useful comments and suggestions.

Author contributions
All authors contributed equally to this work. All authors read and approved the final manuscript.

Funding
This research received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in the Binance repository, https://data.binance.vision.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Received: 16 June 2022 Accepted: 3 November 2022
Published online: 15 January 2023

References

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3):803–821
Barbu VS, Limnios N (2009) Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis. In: Lecture notes in statistics. Springer, New York. https://doi.org/10.1007/978-0-387-73173-5
Bariviera AF, Basgall MJ, Hasperué W, Naiouf M (2017) Some stylized facts of the Bitcoin market. Phys A Stat Mech Appl 484:82–90
Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78
Bouveyron C, Celeux G, Murphy TB, Raftery AE (2019) Model-based clustering and classification for data science: with applications in R. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge
Çinlar E (1975) Markov renewal theory: a survey. Manag Sci 21(7):727–752
D'Amico G (2011) Age-usage semi-Markov models. Appl Math Model 35(9):4354–4366
D'Amico G, Petroni F (2011) A semi-Markov model with memory for price changes. J Stat Mech 12:P12009
D'Amico G, Petroni F (2012a) A semi-Markov model for price returns. Phys A Stat Mech Appl 391(20):4867–4876
D'Amico G, Petroni F (2012b) Weighted-indexed semi-Markov models for modeling financial returns. J Stat Mech 07:P07015
D'Amico G, Petroni F (2018) Copula based multivariate semi-Markov models with applications in high-frequency finance. Eur J Oper Res 267(2):765–777
D'Amico G, Petroni F (2021) A micro-to-macro approach to returns, volumes and waiting times. Appl Stoch Models Bus Ind 37(4):767–789
D'Amico G, Petroni F, Prattico F (2013) First and second order semi-Markov chains for wind speed modeling. Phys A Stat Mech Appl 392(5):1194–1201
D'Amico G, Gismondi F, Petroni F (2018) A new approach to the modeling of financial volumes. In: Silvestrov S, Malyarenko A, Rančić M (eds) Stochastic processes and applications. Springer proceedings in mathematics & statistics. Springer, Cham, pp 363–373
D'Amico G, Lika A, Petroni F (2019) Change point dynamics for financial data: an indexed Markov chain approach. Ann Finance 15(2):247–266
D'Amico G, Masala G, Petroni F, Sobolewski RA (2020a) Managing wind power generation via indexed semi-Markov model and copula. Energies 13(16):4246
D'Amico G, Di Basilio B, Petroni F (2020b) A semi-Markovian approach to drawdown-based measures. Adv Complex Syst 23(08):2050020
De Blasis R (2020) The price leadership share: a new measure of price discovery in financial markets. Ann Finance 16(3):381–405
De Blasis R, Webb A (2022) Arbitrage, contract design, and market structure in Bitcoin futures markets. J Futures Mark 42:492–524
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Stat Methodol 39(1):1–38
Fodra P, Pham H (2015) Semi-Markov model for market microstructure. Appl Math Finance 22(3):261–295
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631
Hafner CM (2020) Testing for bubbles in cryptocurrencies with time-varying volatility. J Financ Econom 18(2):233–249
Janssen J, Manca R (2006) Applied semi-Markov processes, 1st edn. Springer, New York. https://doi.org/10.1007/0-387-29548-8
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429
Levy P (1954) Processus semi-Markoviens. In: Proceedings of the international congress of mathematicians', vol III, North-Holland Publishing Co., Amsterdam, 1956, Amsterdam, pp 416–426
Li T, Kou G, Peng Y, Yu PS (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3109066
Limnios N, Opri¢an G (2003) Ch. 14. An introduction to semi-Markov processes with application to reliability. In: Handbook of statistics, vol 21 of Stochastic processes: modelling and simulation. Elsevier, pp 515–556
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Volume 1: statistics', vol 5.1, University of California Press, pp 281–298. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992
Pasricha P, Selvamuthu D, D'Amico G, Manca R (2020) Portfolio optimization of credit risky bonds: a semi-Markov process approach. Financ Innov 6(1):25
Pyke R (1961a) Markov renewal processes: definitions and preliminary properties. Ann Math Stat 32(4):1231–1242
Pyke R (1961b) Markov renewal processes with finitely many states. Ann Math Stat 32(4):1243–1259
Scott AJ, Symons MJ (1971) Clustering methods based on likelihood ratio criteria. Biometrics 27(2):387–397

Sculley D, (2010) Web-scale k-means clustering. In: Proceedings of the 19th international conference on world wide web. WWW 10. Association for Computing Machinery, New York, pp 1177–1178
Sebestyen G (1962) Decision-making processes in pattern recognition. ACM monograph series, Macmillan. https://books.google.it/books?id=RGZgAAAAMAAJ
Smith WL (1955) Regenerative stochastic processes. Proc Math Phys Eng Sci Proc R Soc A Math Phys 232(1188):6–31
Steinley D (2006) K-means clustering: a half-century synthesis. Br J Math Stat Psychol 59(Pt 1):1–34
Swishchuk A, Hofmeister T, Cera K, Schmidt J (2017) General semi-Markov model for limit order books. Int J Theor Appl Finance 20(03):1750019
Tan S-K, Chan JS-K, Ng K-H (2020) On the speculative nature of cryptocurrencies: a study on Garman and Klass volatility measure. Finance Res Lett 32:101075
Vasileiou A, Vassiliou P-CG (2006) An inhomogeneous semi-Markov model for the term structure of credit risk spreads. Adv Appl Probab 38(1):171–198
Wolfe J (1963) Object cluster analysis of social areas. University of California. https://books.google.it/books?id=RFUdHwAACAAJ

Publisher's Note