European Journal of Operational Research 267 (2018) 765–777

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Copula based multivariate semi-Markov models with applications in high-frequency finance

Guglielmo D'Amico [a], Filippo Petroni [b,*]

[a] Dipartimento di Farmacia, Università G. D'Annunzio, Chieti 66013, Italy
[b] Dipartimento di Scienze Economiche ed Aziendali, Università di Cagliari, Cagliari 09123, Italy

ARTICLE INFO

Article history:
Received 12 August 2015
Accepted 7 December 2017
Available online 14 December 2017

Keywords:
Finance
Stochastic processes
Applied probability
Portfolio analysis

ABSTRACT

We introduce a new multivariate model of multiple asset returns. Our model is based on weighted indexed semi-Markov chains to describe the single (marginals) asset returns, whereas the dependence structure among the considered assets is described by introducing copula functions. A real application of the proposed multivariate model is presented based on the evolution of 6 stocks from the Italian Stock Exchange. We provide empirical evidence that the model is able to correctly reproduce statistical regularities of multivariate real data such as the cross-correlation function, value-at-risk, marginal value-at-risk and conditional value-at-risk. The model is also used for volatility forecasting of each stock.

1. Introduction

High-frequency finance has become one of the most active fields of research in modern finance. From a theoretical standpoint, researchers are interested in understanding financial markets using the entire spectrum of accessible information that, today, is available at very short time scales with price movements at the tick by-tick level. The relevance of high-frequency finance is also driven by practical reasons since practitioners frequently utilize high-frequency data and refuse the use of regular data obtained by transformation and aggregation of the original data. An excellent treatment of high-frequency financial problems and applications is given in Dacorogna, Genay, Muller, Olsen, and Pictet (2001).

Although the literature on modeling of financial returns nowadays is abundant, it is possible to identify two main research lines: the micro-to-macro approach of modeling tick-by-tick asset prices and the econometric or macro-to-micro approach (for more information and comparisons on the two approaches, see Fodra & Pham (2015)). The micro-to-macro approach relies entirely on observable quantities (returns) and therefore is philosophically in contrast with the econometric approach that assumes the observed price to be a collateral effect of an unobservable volatility process to which a noise process transformation is applied (see e.g. Bollerslev 1986). Lattice based models, which include popular binomial and trinomial models (see e.g. Cox, Ross, and Rubinstein (1979) and Boyle (1986)), belong to the micro-to-macro approach. This approach has been further developed in recent years especially by econophysicists. Their contribution relies mainly on the introduction of Continuous Time Random Walks (CTRW) to mimic high-frequency financial data (see e.g. Scalas, Gorenflo, and Mainardi (2000), Mainardi, Raberto, Gorenflo, and Scalas (2000), and Gorenflo, Mainardi, Scalas, and Raberto (2001)). CTRW enables considering waiting times between two consecutive returns as random variables, possibly non-exponentially distributed, which determines non-Markovian behaviors of the return dynamics. Moreover, it is well known that CTRW can be employed to reproduce ordinary diffusions, Levy flights, fractional Brownian motion and ambivalent processes, see Metzler and Klafter (2000). A more complete and general mathematical apparatus is represented by Semi-Markovian Processes (SMP) that were applied in the description of financial returns by D'Amico and Petroni (2012a, 2012b) and subsequently by Fodra and Pham (2015). A SMP allows for arbitrary distributions for the jump size (of the returns in financial applications) and for the waiting time distributions, i.e. the time distribution between two consecutive jumps (changes of value of the return process in the financial framework). SMP based models have been successfully used to reproduce important statistical characteristics of asset prices, such as mean reversion of the price returns and the probability distribution function of the return process. However, D'Amico and Petroni (2012a, 2012b) showed that SMP were unable to correctly reproduce the autocorrelation of squared returns at 1 minute frequency. The solution to this important problem was achieved through the advancement of a new class of stochastic models called weighted-indexed semi-Markov chains (WISMC), see

* Corresponding author.
E-mail address: fpetroni@unica.it (F. Petroni).

https://doi.org/10.1016/j.ejor.2017.12.016
0377-2217/

D'Amico and Petroni (2011, 2012b). In these papers the authors have demonstrated that the WISMC model is able to reproduce the statistical properties of one-dimensional return financial data and, in particular, the clustering of volatility can be efficiently described modeling only the returns without a volatility model as in the econometric approach. The WISMC model provides a very general approach that encompasses both SMP and CTRW. The superiority of the WISMC model resides in its accurate probabilistic description of the asset return evolution, which accounts for the serial dependence of asset returns by incorporating past events (trade times and return sizes) through an index process that increases the memory of the process. In this respect, WISMC incorporates the idea of a self-exciting process into the semi-Markov framework, as advanced by Hawkes (1971) for point processes and further extended in Hautsch (2004), Bowsher (2007), and Bauwens and Hautsch (2009).

Another important topic in modern finance is the study of the dependence structure between a pool of assets and the calculated risk of a portfolio (see e.g. Breymann, Dias, & Embrechts (2003)). This factor is crucial during asset allocation for diversification of a portfolio. The traditional approach is to measure the correlation coefficients among the assets, however since the correlation coefficient is only a measure of linear dependence, better results can be obtained by exhaustively quantifying the dependence structure. The copula function revealed to be a useful method to model the dependence structure among financial data, e.g. Cherubini, Luciano, and Vecchiato (2004). Indeed, the use of copulas together with the marginal distributions of each financial asset return permits the construction of the joint cumulative distribution function of a vector of financial asset returns. Copula based models are very versatile and have been applied extensively in modern finance (see e.g. Embrechts, Hoing, and Juri (2003), Chan and Kroese (2010), Durante and Jaworski (2010), Sak, Hormann, and Leydold (2010), Laih (2014), Kakouris and Rustem (2014), and Durante, Fernández-Sánchez, Quesada-Molina, and Úbeda-Flores (2015)).

In this paper, we aim to provide a new and tractable multivariate model to describe the evolution of a vector of asset returns. The marginal models are assumed to follow WISMC models and the dependence structure among the return processes is embodied by using copula functions. Our goal is to extend WISMC models into a multivariate setting by defining a WISMC-copula model to show that the dependence structure can also be opportunely studied with this class of stochastic processes. More precisely we demonstrate that the advanced multivariate model, once estimated by using real data, is able to reproduce the probability mass function of all the stock returns considered in our portfolio, the autocorrelation function of each stock, and the dependence structure between the considered stocks and therefore can also be used to determine portfolio risk measures such as the Value-at-Risk (VaR), the Marginal Value-at-Risk and the Conditional Value-at-Risk (C-VaR). Finally, the model is used for volatility forecasting of each stock and the obtained results are compared with those obtained using GARCH models. The WISMC-copula proves to be a valuable model for studying high-frequency return dynamics as demonstrated by the notable aforementioned empirical results. Indeed, a stochastic model of asset evolution should be able to correctly reproduce in-sample characteristic of returns, in order to be a realistic model, but it should also be efficient when used for out-of-sample analysis, which is relevant for efficient asset allocation in a financial portfolio. The WISMC-copula model has several important practical consequences, just to name a few, it could be used by practitioners in order to understand the volatility of financial assets that, in turn, help to quantify the risk of an investment and also to optimize the portfolio for optimal allocation of wealth. The assessment of VaR and C-VaR allows for a realistic portfolio optimization based on mean-VaR framework, see Lwin, Qu, and MacCarthy (2017) and on linear programming models, see Mansini, Ogryczak, and Speranza (2007).

The application of the WISMC-copula model necessitates the use of different operational research techniques such as stochastic processes, Monte Carlo simulations and optimization techniques aimed to estimate the memory parameter of the model. For this reason, we think that this paper could be of relevance to both financial and operational research communities.

The rest of the paper is organized as follows: Section 2 presents the marginal WISMC model and the copula based multivariate model. Section 3 discusses the data and the empirical results of the model are compared with those obtained on real data. Section 4 presents some concluding remarks. Appendix 1 summarizes a table of notation. Appendix 2 provides empirical estimators of the weighted-indexed semi-Markov kernel obtained by extending the approach advanced in Barbu and Limnios (2006) for ordinary semi-Markov chains. Appendix 3 illustrates an algorithm describing the different steps that are necessary for the application of the model.

2. Mathematical models

The weighted-indexed semi-Markov chain (WISMC) model is first described for the marginal distribution of each asset return. Then, a step by step procedure is introduced for the multivariate distribution of multiple asset returns. The dependence structure among the assets is described by using copulas.

2.1. Marginals weighted-indexed semi-Markov model

The WISMC model enables efficiently reproducing long-term dependence on stock returns, see D'Amico and Petroni (2011, 2012a, 2012b).

Let $S(t)$ be the price of a financial asset at time $t \in \mathbb{N}$. The time varying log returns, defined as $R(t) = \log(S(t)/S(t-1))$, are transformed into a series of discrete returns by means of a map:

$$\mathcal{M} : \mathbb{R} \longrightarrow E = \{-z_{\min}\Delta, \ldots, -\Delta, 0, \Delta, \ldots, z_{\max}\Delta\},$$

where Δ is the grid amplitude of E.

We assume that the discrete return $R_d(t) := \mathcal{M}(R(t))$ attains the value $i\Delta$ whenever the continuous return $R(t)$ belongs to the interval $\left((i-\frac{1}{2})\Delta, (i+\frac{1}{2})\Delta\right]$. The lowest discrete return $R_d(t) = -z_{\min}\Delta$ is achieved when the continuous return $R(t) \leq (-z_{\min} + \frac{1}{2})\Delta$ whereas the highest discrete return $R_d(t) = z_{\max}\Delta$ is assigned whenever $R(t) > (z_{\max} - \frac{1}{2})\Delta$.

The sequence of discrete returns $\{R_d(t)\}_{t \in \mathbb{N}}$ is successively converted into a series of returns $\{J_n\}_{n \in \mathbb{N}}$ and corresponding jump times $\{T_n\}_{n \in \mathbb{N}}$ of the asset returns. To this purpose it is sufficient to set $T_0 = 0$, $J_0 = R_d(0)$ and for $n \geq 1$

$$T_n = \inf\{t \in \mathbb{N}, t > T_{n-1} : R_d(t) \neq R_d(T_{n-1})\}, \tag{1}$$

$$J_n = R_d(T_n). \tag{2}$$

We also consider an additional variable $\{V_n^\lambda\}_{n \in \mathbb{N}}$ that describes the value of an index process at the nth jump time. According to D'Amico and Petroni (2012b) we set

$$V_n^\lambda = \sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} f^\lambda(J_{n-1-k}, T_n, a), \tag{3}$$

where f is any real value bounded function and V_0^λ is known and non-random. The variable V_n^λ is designated to summarize the information contained in the past trajectory of the return process that

should be used in order to increase the memory of the process. Indeed, to each past discrete return J_{n-1-k} occurred at time $a \in \mathbb{N}$ is associated the value $f^{\lambda}(J_{n-1-k}, T_n, a)$ that also depends on the current time T_n. The quantity λ denotes a parameter that represents a weight and should be calibrated to data. In the applicative section we will describe the calibration of λ as well as the choice of the function f.

The WISMC model is specified once a dependence structure between the variables is considered. Toward this end, the following assumption is formulated:

$$\begin{aligned}\mathbb{P}[J_{n+1} &= j,\ T_{n+1} - T_n \le t | J_n, T_n, V_n^{\lambda}, J_{n-1}, T_{n-1}, V_{n-1}^{\lambda}, \ldots] \\ &= \mathbb{P}[J_{n+1} = j,\ T_{n+1} - T_n \le t | J_n, V_n^{\lambda}] := Q_{J_n j}^{\lambda}(V_n^{\lambda}; t). \end{aligned} \tag{4}$$

Relation (4) asserts that the knowledge of the values of the variables J_n, V_n^{λ} is sufficient to give the conditional distribution of the couple $J_{n+1}, T_{n+1} - T_n$ whatever the values of the past variables might be. Therefore, to compute joint probabilities of return and transition times, we need the knowledge of the last state of return and the last value of the index process.

The matrix of functions $\mathbf{Q}^{\lambda}(v;t) = (Q_{ij}^{\lambda}(v;t))_{i,j \in E}$ is called the *weighted-indexed semi-Markov kernel*. If $\mathbf{Q}^{\lambda}(v;t)$ is constant in v then the WISMC kernel degenerates in an ordinary semi-Markov kernel and the WISMC model becomes equivalent to classical semi-Markov chain model, see e.g. D'Amico and Petroni (2012a) and Fodra and Pham (2015).

The three-dimensional process $\{J_n, T_n, V_n^{\lambda}\}$ describes the system in correspondence of any jump time T_n. However, in order to define the multivariate model, it is important to describe the system in correspondence of any time t, which can be a jump time ($t = T_n$) or not ($t \neq T_n$). To this end, let us also define $N(t) = \sup\{n \in \mathbb{N} : T_n \le t\}$ as the process that counts the number of jumps up to time t. The random process $Z(t) := J_{N(t)}$ is called weighted-indexed semi-Markov chain. It denotes the state of the system (price return) at time t. Now the index process can be extended to any time $t \in \mathbb{N}$ as follows:

$$V^{\lambda}(t) = \sum_{k=0}^{N(t)-1+\theta} \sum_{a=T_{N(t)+\theta-k}}^{(t \wedge T_{N(t)+\theta-k})-1} f^{\lambda}(J_{N(t)+\theta-1-k}, t, a), \tag{5}$$

where $\theta = 1_{\{t > T_{N(t)}\}}$. If $t = T_n$, then t is a jump time, we have that $V^{\lambda}(t) = V_n^{\lambda}$.

2.2. The multivariate weighted-indexed semi-Markov model

In this section we extend the WISMC model into a multivariate setting. Suppose that we have m stocks. We aim to specify a model such that the joint return distribution is formulated so that the marginals analysis yields model compatible with a WISMC model $Z_h^{\lambda_h}(t)$ of kernel $\mathbf{Q}_h^{\lambda_h}(v;t)$, $h = 1, 2, \ldots, m$.

Copula models are suitable for this purpose. The construction of the multivariate model passes through a step by step procedure, which is an extension of that advanced by van Boxtel (2007) for a simple credit risk model based on Markov chains. More precisely, the WISMC model collapses into a Markov chain model whenever, for all values v of the index process, we have

$$Q_{ij}^{\lambda}(v;t) = p_{ij} 1_{\{t=1\}}.$$

In this simplified case, transition probability functions are independent of past information incorporated by the index process and are non-zero only for a waiting time $t = 1$, i.e. there is not randomness in the sequence of trade times that have exactly one time unit in length. The Markov chain hypotheses are in contrast with the empirical evidence of high-frequency finance, as demonstrated in D'Amico and Petroni (2011).

The first step of the procedure requires representing the transition probability function of the WISMC model over a time horizon of one period. This problem was solved in D'Amico and Petroni (2014) by extending, to the WISMC framework, a technique first proposed in Vassiliou and Papadopoulou (1992) for classical semi-Markov chains. The technique introduces the backward recurrence time process $B(t) := t - T_{N(t)}$ and describes the probabilistic behavior of the Markov process $(Z(t), B(t))$ on the extended state space $E \times \mathcal{D}$ where $\mathcal{D} = \{0, 1, \ldots, D\}$ and D is the maximum length of stay in the states of the process.

Let denote by

$$\begin{aligned} p_{(i,u,v)}(j,d) :&= \mathbb{P}[Z^{\lambda}(n+1) = j, B(n+1) = d \mid Z^{\lambda}(n) = i, \\ &\quad B(n) = u, V^{\lambda}(n) = v], \end{aligned} \tag{6}$$

the one-step transition probability function of the WISMC model. First, it should be noted that $d \in \{0, u+1\}$ because if a transition from state i to an arbitrary state $j \neq i$ is executed, then

$$B(n+1) = n + 1 - T_{N(n+1)} = n + 1 - (n+1) = 0.$$

On the contrary, if the system will remain in state i next period, then, being $T_{N(n+1)} = T_{N(n)}$, we have

$$B(n+1) = n + 1 - T_{N(n+1)} = 1 + (n - T_{N(n)}) = 1 - B(n) = 1 + u.$$

As demonstrated in D'Amico and Petroni (2014), the probability (6) can be obtained as follows:

$$p_{((i,u,v)}(j,d) = \begin{cases} \dfrac{\bar{H}_i^{\lambda}(v + \Delta V(N(n), n); 1+u)}{\bar{H}_i^{\lambda}(v + \Delta V(N(n), n); u)} & \text{if } j = i,\ d = 1+u \\ \dfrac{q_{ij}^{\lambda}(v + \Delta V(N(n), n); 1+u)}{\bar{H}_i^{\lambda}(v + \Delta V(N(n), n); u)} & \text{if } j \neq i,\ d = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\bar{H}_i^{\lambda}(v;t) = 1 - \sum_{j \in E} Q_{ij}^{\lambda}(v;t)$ is the survival function of sojourn time in state i, $q_{ij}^{\lambda}(x,t) = Q_{ij}^{\lambda}(x,t) - Q_{ij}^{\lambda}(x,t-1)$ and $\Delta V(N(n), n) = V_{N(n)}^{\lambda} - V^{\lambda}(n)$ is the opposite of the increment of the index process on the waiting time $n - N(n)$.

If we know the value $V^{\lambda}(n) = v$, and $T_{N(n)} = n - u$ with $n \ge u \ge 0$ and $T_{N(n)+1} > n$, then

$$\begin{aligned} V_{N(n)}^{\lambda} = v &- \sum_{a=T_{N(n)}}^{n-1} f^{\lambda}(J_{N(n)}, n, a) \\ &+ \sum_{k=0}^{N(n)-1} \sum_{a=T_{N(n)-1-k}}^{T_{N(n)-k}-1} \Delta f^{\lambda}(J_{N(n)-k}, T_{N(n)}, n, a), \end{aligned} \tag{8}$$

where $\Delta f^{\lambda}(i, T_{N(n)}, n, a) := f^{\lambda}(i, T_{N(n)}, a) - f^{\lambda}(i, n, a)$. Therefore relation (8) allows the computation of the increment of the index process that is necessary in the calculation of the transition probability (7).

As a second step of the procedure we compute the cumulative distribution function (cdf) of the asset return and we extend it into a continuous setting.

For all $i_h \in E$, $u_h \in \mathcal{D}$ and $v_h \in \mathbb{R}$ let us define the cumulative distribution function $P_{(i_h,u_h,v_h)}(\cdot,\cdot)$:

$$\begin{aligned} P_{(i_h,u_h,v_h)}(j_h,d_h) &= \mathbb{P}[Z_h^{\lambda}(n+1) \le j_h, B_h^{\lambda}(n+1) \le d_h \mid Z_h^{\lambda}(n) = i_h, \\ &\quad B_h^{\lambda}(n) = u_h, V_h^{\lambda}(n) = v_h] \end{aligned} \tag{9}$$

$$= \begin{cases} \sum_{\substack{s < j_h \\ s \neq i_h}} p_{(i_h,u_h,v_h)}(s,0) & \text{if } i_h > j_h,\ i_h, j_h \in E \\ \sum_{s \le j_h} p_{(i_h,u_h,v_h)}(s,0) & \\ \quad + p_{(i_h,u_h,v_h)}(i_h, u_h+1) 1_{\{d_h \ge u_h+1\}} & \text{if } i_h \le j_h,\ i_h, j_h \in E. \end{cases}$$

For any triple (i_h, u_h, v_h) we define the set of admissible states

$$\mathcal{A}_{(i_h,u_h,v_h)} = \{(j_h, d_h) \in E \times \mathcal{D} : p_{(i_h,u_h,v_h)}(j_h, d_h) \neq 0\}. \quad (10)$$

An admissible state is maximal if and only if $j_h = z_{\max}$.

Let us consider any non maximal state $(r_h, d_h) \in \{(-z_{\min}, \cdot)\} \times \mathcal{A}_{(i_h,u_h,v_h)}$ and denote by

$$s_h = \inf\{j_h : (j_h, d_h) \in \mathcal{A}_{(i_h,u_h,v_h)}, \forall d_h \in \mathcal{D}\},$$

with the convention that $\inf\{\emptyset\} = z_{\max}$. The state s_h denotes the successive adjacent admissible state to (r_h, d_h).

Now it is possible to extend the cdf (9) into a continuous one by using, for example, a simple linear interpolation map. To this end, for a non integer x, with $r_h < x < s_h$ we define:

$$P_{(i_h,u_h,v_h)}(x, d_h) =$$

$$\begin{cases} P_{(i_h,u_h,v_h)}(r_h, 0) + (x - r_h)\Big(P_{(i_h,u_h,v_h)}(s_h, 0) - P_{(i_h,u_h,v_h)}(r_h, 0)\Big) & \text{if } i_h < r_h \text{ or } r_h < i_h < s_h \text{ or } i_h > s_h \\ P_{(r_h,u_h,v_h)}(r_h, u_h + 1) + (x - r_h)\Big(P_{(r_h,u_h,v_h)}(s_h, 0) - P_{(r_h,u_h,v_h)}(r_h, u_h + 1)\Big) & \text{if } i_h = r_h \\ P_{(s_h,u_h,v_h)}(r_h, 0) + (x - r_h)\Big(P_{(s_h,u_h,v_h)}(s_h, u_h + 1) - P_{(s_h,u_h,v_h)}(r_h, 0)\Big) & \text{if } i_h = s_h \end{cases}$$

The function $P_{(i_h,u_h,v_h)}(\cdot, \cdot)$ can be thought like the cdf of a continuous return index $X_h^\lambda(n)$ of the stock h at any time n:

$$P_{(i_h,u_h,v_h)}(x_h, d_h) = \mathbb{P}[X_h^\lambda(n+1) \leq x_h, B_h^\lambda(n+1) \leq d_h \mid Z_h^\lambda(n) = i_h, B_h^\lambda(n) = u_h, V_h^\lambda(n) = v_h].$$

The continuous cdf is by construction strictly increasing except in the interval $[\max\{\mathcal{A}_{(i_h,u_h,v_h)}\}, z_{\max}]$ where the cdf is constant and equal to 1. It should be noted that if $z_{\max} \in \mathcal{A}_{(i_h,u_h,v_h)}$, then, the cdf is strictly increasing on its support $[-z_{\min}, z_{\max}]$. Let $P^{-1}_{(i_h,u_h,v_h)}(\cdot) : [0, 1] \to \mathbb{R}$ be the quantile function associated with the cdf $P_{(i_h,u_h,v_h)}(\cdot, \cdot)$. Then

$$P^{-1}_{(i_h,u_h,v_h)}(\alpha) = \inf\{x : P_{(i_h,u_h,v_h)}(x, 0) \geq \alpha\},$$

for every $\alpha \in [0, 1]$. The quantile function, for a given value $\alpha \in [0, 1]$, gives the successive continuous state of the asset return $X_h^\lambda(n+1)$.

The successive discrete return $Z_h^\lambda(n+1)$ is determined by rounding up to the adjacent discrete return state, i.e.

$$(Z_h^\lambda(n+1), B_h(n+1)) = \begin{cases} (j_h, 0) & \text{if } \lceil X_h^\lambda(n+1) \rceil = j_h \neq i_h \\ (i_h, B_h^\lambda(n) + 1) & \text{if } \lceil X_h^\lambda(n+1) \rceil = i_h. \end{cases}$$

The probability distribution of the successive discrete return can be expressed by using the quantile function and a standard Uniform random variable $U_{h,n}$.

The third step of the procedure makes use of copula functions to get the joint distribution of asset returns. Copula functions can be used for the purpose of building models with dependent components having fixed marginals, see e.g. Joe (1997) and Nelsen (2006).

The aim is to specify a model such that: the return dynamic is formulated so that the marginal analysis of each asset process yields parameters compatible with a WISMC model, while, at the same time, there is a dependence among the considered m assets. The joint distribution of the financial returns $(X_1^\lambda(n+1), X_2^\lambda(n+1), \ldots X_m^\lambda(n+1))$ is defined conditional on the discrete returns at time n denoted by the vector $\mathbf{s}(n) = (s_1(n), \ldots, s_m(n))$, on the values of index process $\mathbf{v}(n) = (v_1(n), \ldots, v_m(n))$ and on the values of backward recurrence time processes $\mathbf{u}(n) = (u_1(n), \ldots, u_m(n))$. For any vector of returns $\mathbf{x}$ and of durations $\mathbf{d}$ at time $n+1$ we define the joint probability function

Table 1
Stocks used in the application and their symbols.

F	Fiat
ISP	Intesa San Paolo
MS	Mediaset
PC	Pirelli
TIT	Telecom
TEN	Tenaris

$$R_{(\mathbf{s}(n),\mathbf{u}(n),\mathbf{v}(n))}(\mathbf{x}, \mathbf{d}) := \mathbb{P}[\mathbf{X}(n+1) \leq \mathbf{x}, \mathbf{B}(n+1) \leq \mathbf{d} \mid \mathbf{X}(n) = \mathbf{s}(n), \mathbf{B}(n) = \mathbf{u}(n), \mathbf{V}(n) = \mathbf{v}(n)].$$

Given that copulas are functions that completely describe the dependence among continuous random variables, the joint distribution of the financial returns can be defined through a copula $\mathcal{C}$ as

$$R_{(\mathbf{s}(n),\mathbf{u}(n),\mathbf{v}(n))}(\mathbf{x}, \mathbf{d}) = \mathcal{C}\Big(P_{(s_1,u_1,v_1)}(x_1, d_1), P_{(s_2,u_2,v_2)}(x_2, d_2), \ldots, P_{(s_m,u_m,v_m)}(x_m, d_m)\Big).$$

Notice that, only for simplicity reasons, we suppose that the dependence structure is constant on time. However, in a general approach, it is possible to consider time-varying dependence structures. This multivariate model permits the simulation of dependent asset returns by drawing dependent uniform values from the copula at any time $n \in \mathbb{N}$.

3. Application

The multivariate model, as described in the previous sections, was applied to a portfolio of 6 stocks from the Italian Stock Exchange ('Borsa Italiana').

The list of stocks analyzed and their symbols are reported in Table 1.

The database is composed of tick-by-tick quotes recorded from January 2007 to December 2010 (4 full years). Data have been resampled to have a 1 minute frequency. The number of returns analyzed is then roughly 500×10^3 for each stock. A better description of the database can be found in D'Amico and Petroni (2011). Returns have been discretized into 5 states (see below for a better explanation of the discretization procedure) chosen to be symmetrical with respect to returns equaling zero and to keep the shape of the distribution unchanged. Returns are in fact already discretized in real data due to the discretization of stock prices, which are fixed by each stock exchange and are dependent on the value of the stock. Just to give an example, in the Italian stock market for stocks with value between 5.0001 and 10 euros, the minimum variation is fixed to 0.005 euros (usually called tick). We then tried to remain as close as possible to this discretization.

In Table 2 we summarize the descriptive statistics of the dataset.

Fig. 1 shows the histograms for the different stocks and a graphical comparison with Gaussian distributions. Following D'Amico and Petroni (2012b) we use, as defined by the function f^λ in (3), an exponentially weighted moving average (EWMA) of the squares of returns, which is expressed as

$$f^\lambda(J_{n-1-k}, T_n, a) = \frac{\lambda^{T_n - a} J_{n-1-k}^2}{\sum_{k=0}^{n-1} \sum_{a=T_{n-k}}^{T_{n-k}-1} \lambda^{T_n - a}} = \frac{\lambda^{T_n - a} J_{n-1-k}^2}{\sum_{a=1}^{T_n} \lambda^a}, \quad (11)$$

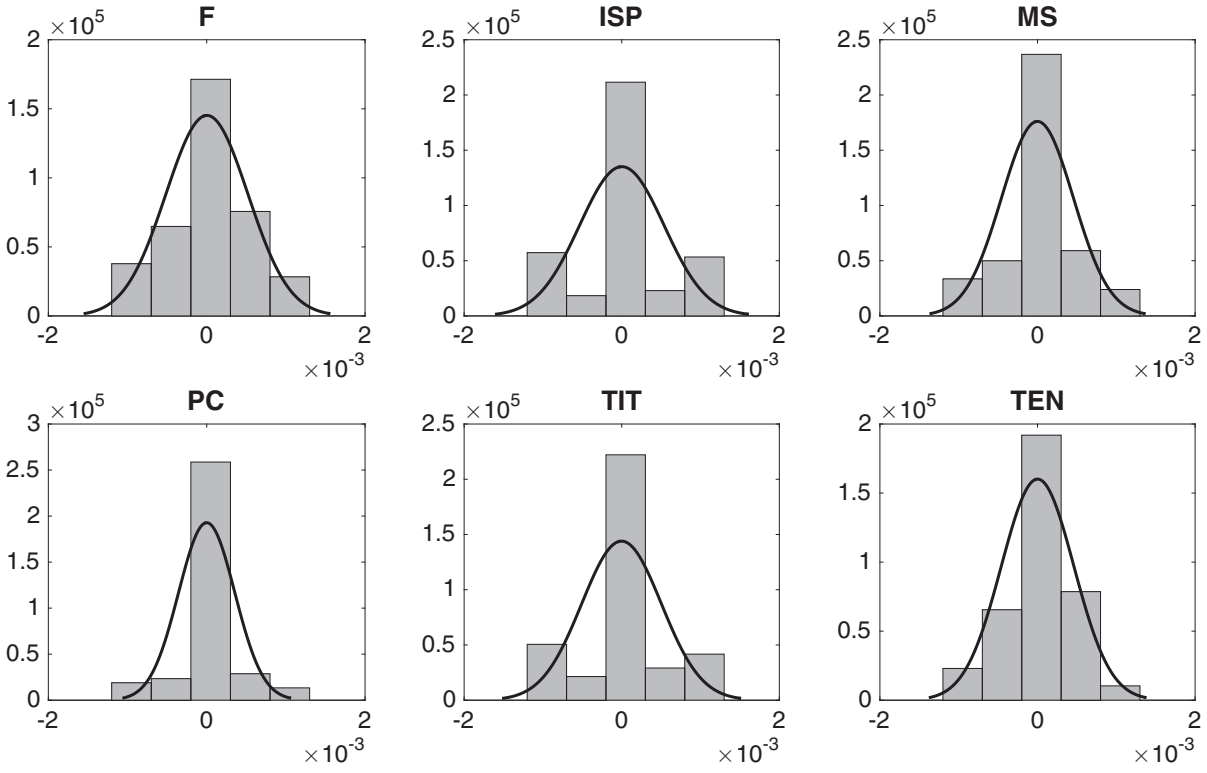

Fig. 1. Histograms of returns of the different stocks compared with Gaussian distributions.

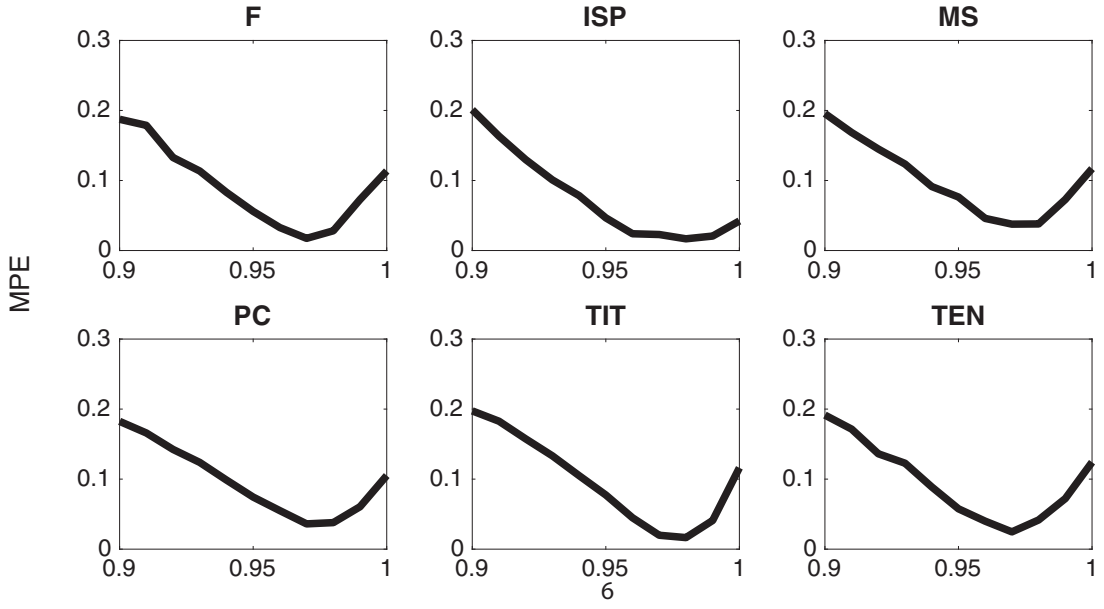

Fig. 2. Optimization of the parameter λ for each stock based on minimization of MPE.

Table 2
Descriptive statistics of the dataset used for the analysis.

Stock	Mean	Median	Standard deviation	Skewness	Kurtosis
F	2.86×10^{-6}	0	0.52×10^{-3}	-0.21×10^{-2}	2.28
ISP	1.26×10^{-6}	0	0.54×10^{-3}	-0.24×10^{-2}	2.62
MS	-7.89×10^{-7}	0	0.46×10^{-3}	0.39×10^{-3}	2.88
PC	-1.97×10^{-7}	0	0.35×10^{-3}	0.37×10^{-2}	4.75
TIT	-2.41×10^{-6}	0	0.50×10^{-3}	0.59×10^{-3}	2.96
TEN	9.81×10^{-7}	0	0.46×10^{-3}	-0.77×10^{-3}	2.39

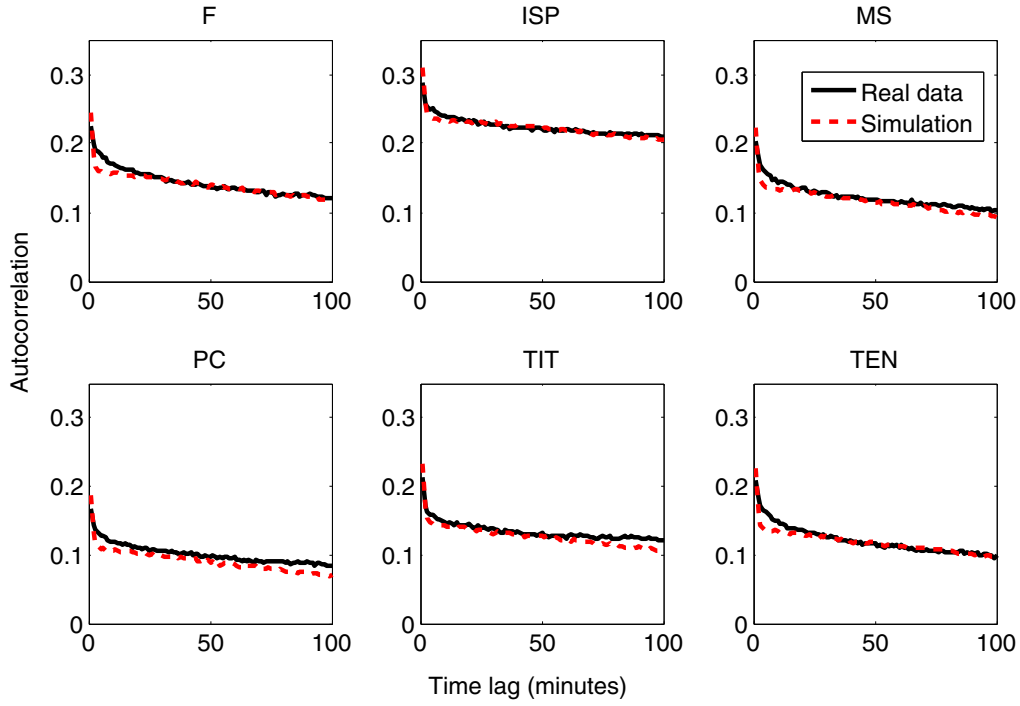

Fig. 3. Autocorrelation functions of the square of returns for real data (solid line) and synthetic (dashed line) time series.

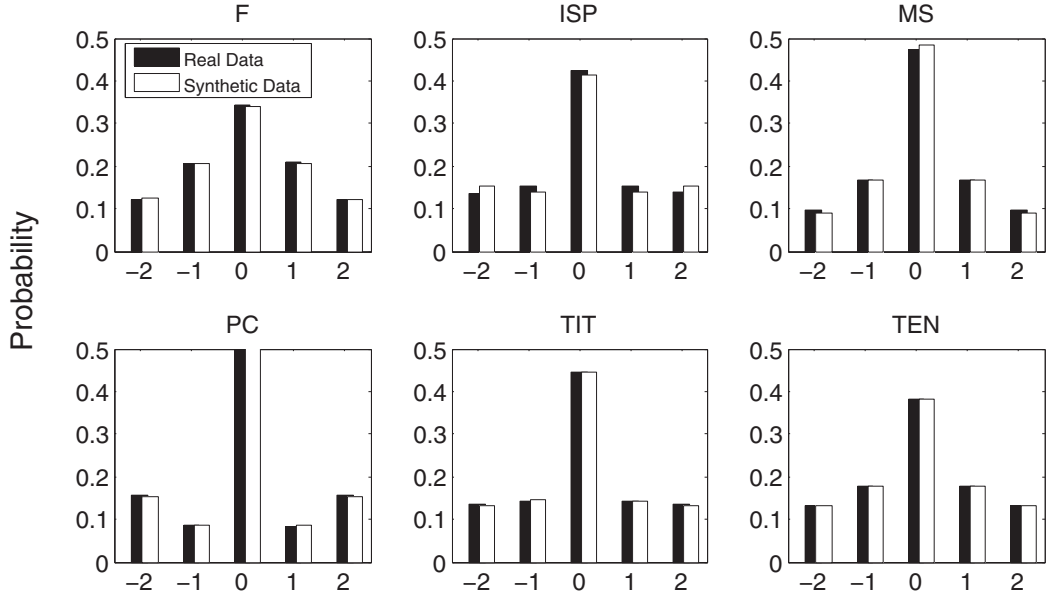

Fig. 4. Comparison of the probability mass function for real and synthetic data for each stock.

and consequently the index process becomes

$$V_n^{\lambda} = \sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \left(\frac{\lambda^{T_n - a} J_{n-1-k}^2}{\sum_{a=1}^{T_n} \lambda^a} \right). \tag{12}$$

The index V^{λ} was also discretized into 5 states of low, medium low, medium, medium high and high volatility. Using these definitions and discretization we estimated, for each stock, the probabilities defined in the previous section by using their estimators directly from real data. Appendix 2 contains the derivations of the empirical estimators of the weighted-indexed semi-Markov kernel. By means of Monte Carlo simulations, we were able to produce, for each of the 6 stocks, a synthetic time series.

To preserve the cross-correlation between stocks we used a Gumbel copula, the parameters of the copula were estimated by maximum likelihood estimators. The choice of the Gumbel copula was made based on the fact that this copula preserves the heavy tails of the distribution of returns. It is important to note at this stage that, in our case and for our database, it would be also possible to use a simple Gaussian copula. In fact, the database used here for the analysis is made of high frequency intraday data that do not show any heavy tails behavior. The frequency of trade is very high and then returns are all limited within a small range (between −0.5% and 0.5%).

Each time series is a realization of the stochastic process described in the previous section with the same time length of real

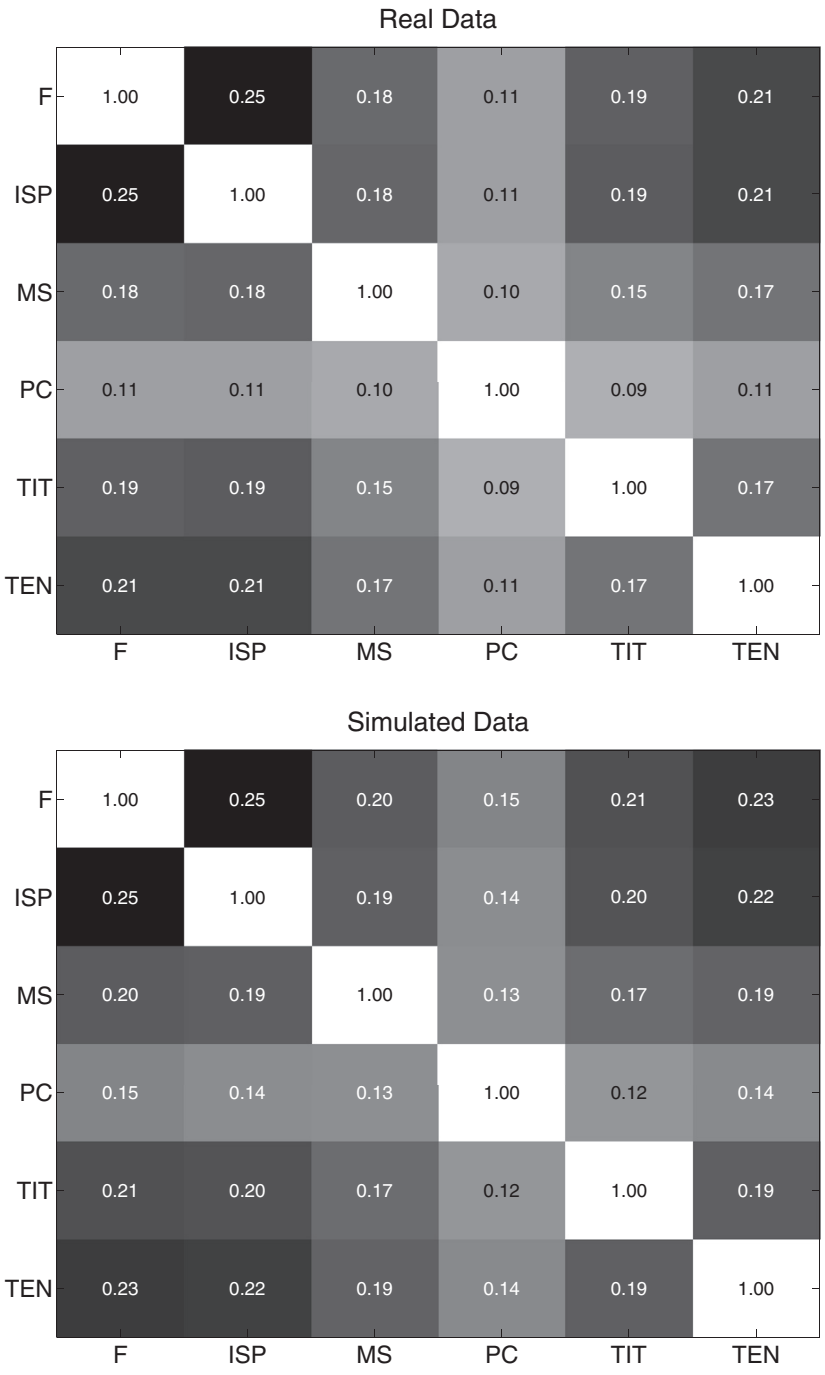

Fig. 5. Cross-correlation matrices for real (up) and synthetic (down) data.

data. Statistical features of these synthetic time series are then compared with the statistical features of real data. In particular, we tested our model for the ability to reproduce the autocorrelation functions of the squares of returns (volatility) and the cross-correlation betweens stocks. Recalling the definition of the autocorrelation function: if R indicates returns, the time lagged (τ) autocorrelation of the square of returns is defined as

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

$\Sigma(\tau)$ was used to optimize the parameters of the model. In particular, by computing the mean percentage error (MPE) between real and simulated $\Sigma(\tau)$, all the parameters of the model were chosen to minimize the MPE.

3.1. Optimization parameters

The model presented here depends mainly on two parameters that have to be optimized with respect to the used database. They are: the number of states for the return process and the value of λ. Both parameters were optimized by minimizing the mean percentage error between the real and synthetic autocorrelation functions. The model was tested by changing the number of states in the discretization of data and, for each discretization, the value of λ was changed within the range [0.9, 1]. First of all, recalling that the database used for this analysis is composed of intra day high frequency data, returns are then very small and within a small range of values (between -0.5% and 0.5%) with almost zero probability

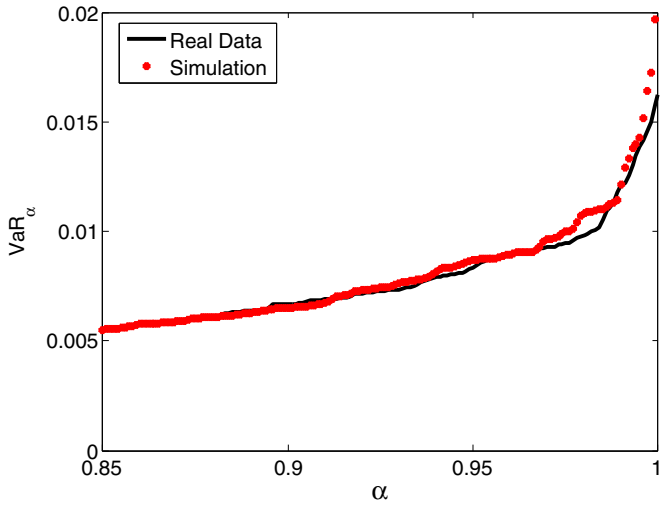

Fig. 6. Comparison of the Value at Risk as a function of the parameter α for real and simulated data. The percentage MPE is 3%.

of falling outside this range (i.e. 99.5% of returns fall within this range). Nevertheless, we ran extensive Monte Carlo simulations to better choose the number of states. We found that statistical features are better described when the number of states increases. We tested the model with 3, 5, 7, 9, 11 equally spaced states ranging in the return distribution. Results show (see Table 3 for the FIAT stock) that the ability of the model to reproduce statistical features of real data (i.e. distribution and autocorrelation function) increases very rapidly between 3 and 5 states, while from 5 to 7 states and over it remains almost constant. This feature is common to all the considered stocks.

Table 3
Mean percentage error (MPE) between real and synthetic autocorrelation function for different number of states and for the best value of λ.

Number of states	MPE (%)
3	10.3
5	2.4
7	2.2
9	2.2
11	2.1

Therefore, we chose 5 states as the better value for the model. This choice was also made by considering a trade off between accuracy of the description of returns distribution and number of parameters to be estimated. An increase in the number of states better describes the process but requires a larger dataset to obtain reliable estimates, and it could also be unnecessary to achieve the accuracy needed in modeling price variation. The number of states and the value of λ are then optimized contemporaneously. In Fig. 2 we show the results for λ optimization with returns discretized into 5 states. It is possible to note that the shape of the MPE is similar for all the considered stocks even if the optimal values for λ (corresponding to the minimum point of each curve) are not the same and depend on the specific stock. The shape of the curve follows an intuitive path: low values of λ produce high MPE because the weight attributed to past observations is too small and relevant past information is not adequately incorporated to explain future dependence. As λ increases, the MPE decreases until a minimum value λ^* that represents the optimal weight. If we consider too large values of $\lambda(>\lambda^*)$ then the MPE rises again because the model places excessive weight on past information that is not relevant for future returns.

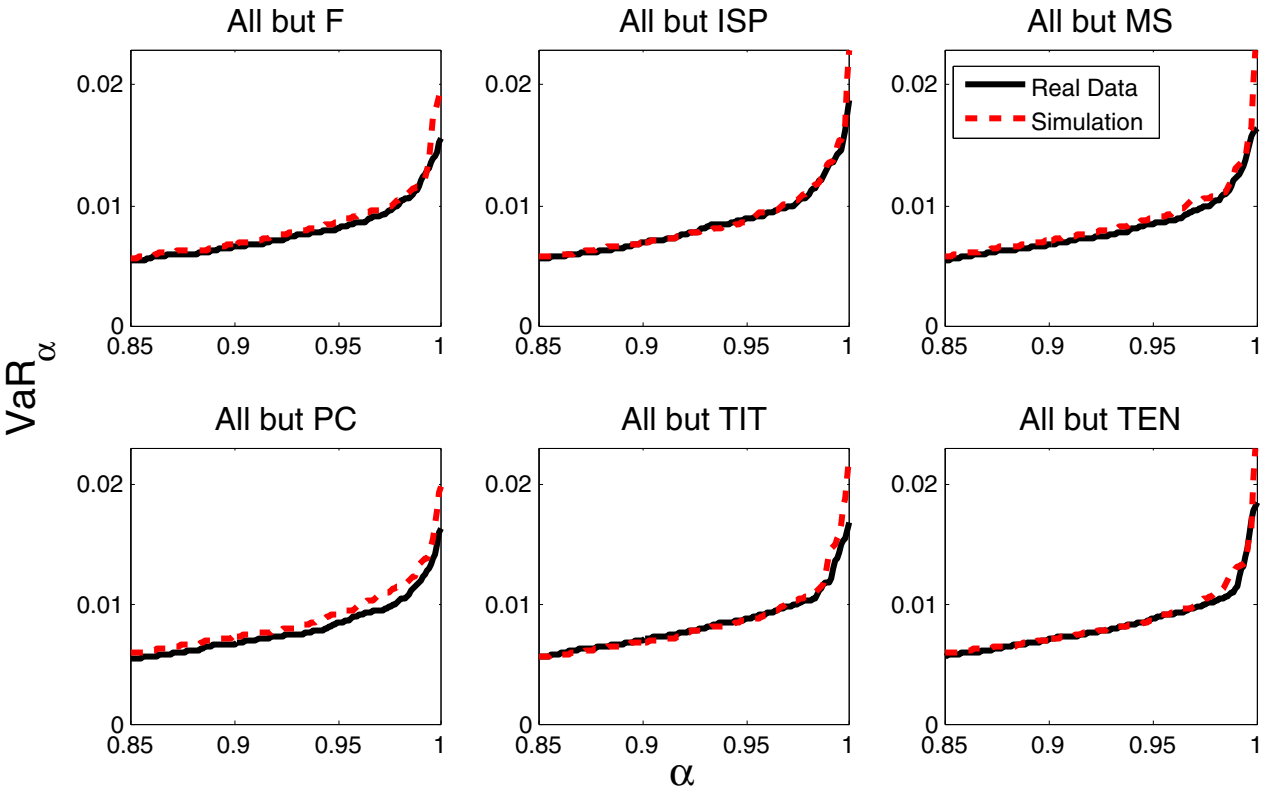

Fig. 7. Marginal value at risk.

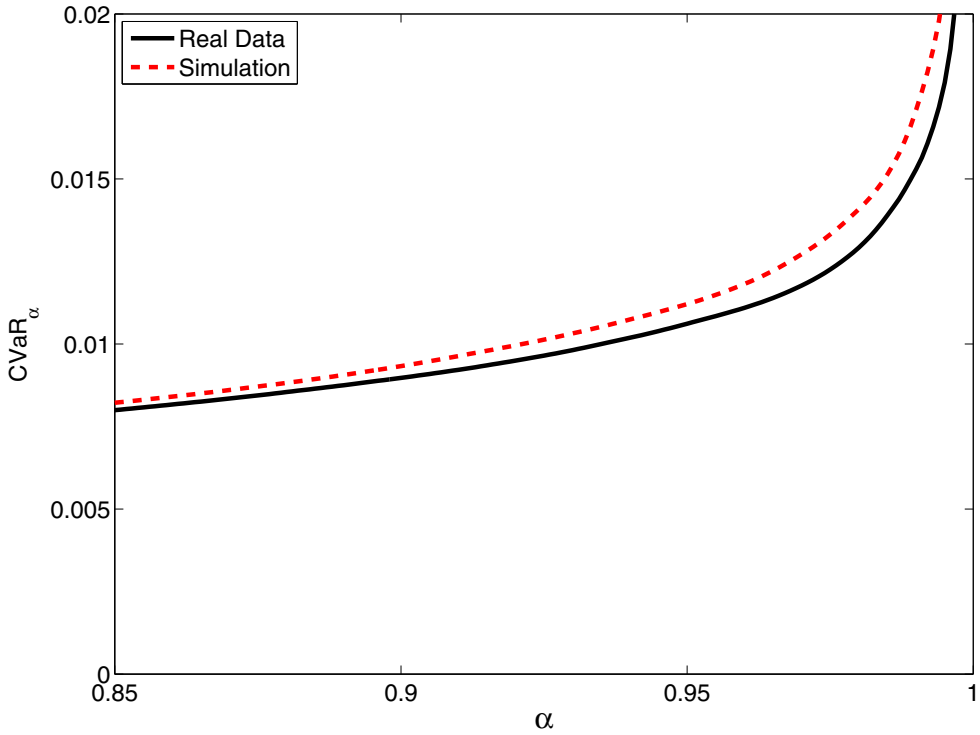

Fig. 8. Comparison of the conditional value-at-risk as a function of the parameter α for real and simulated data. The MPE is 5%.

Table 4
Mean percentage error (MPE) between real and synthetic autocorrelation function reported for WISMC model and compared with different GARCH models. The * denotes the best as suggested by AIC and BIC criterion.

Stock	WISMC (%)	GARCH(1,1) (%)	GARCH(1,2) (%)	GARCH(2,1) (%)
F	2.4	8.5	5.6	7.1*
ISP	2.1	2.6*	5.6	10.1
MS	2.8	7.4	6.7*	7.8
PC	3.2	8.1	9.3	8.4*
TIT	2.6	6.4	3.4	9.6*
TEN	2.1	2.7	5.4	6.3*

3.2. Results on statistical features

Once the parameters are optimized, it is possible to compare results between real and synthetic data. In Fig. 3 we compare $\Sigma(\tau)$ for real data and for synthetic data modeled by using the best parameters for each of the analyzed stock.

For comparison reasons, we also demonstrate the mean percentage error obtained by using three different GARCH models. The parameters of the GARCH models were obtained by applying a maximum likelihood optimization algorithm using Matlab software, see results in Table 4. It is possible to note that our model is able to almost perfectly reproduce the volatility autocorrelation of these stocks and it performs better, at least in regards to the autocorrelation function, than the chosen GARCH models. In Table 4 the symbol * denotes the best GARCH model selected according to both Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). It is possible to see that the selected GARCH is not the same for all stocks and that AIC and BIC do not always choose the model that best reproduces the autocorrelation function.

In the calculation we used GARCH models with Gaussian noise. This choice is supported by real data of the considered stocks. Indeed, we applied Engle's ARCH test (see Engle (1988)) on the residuals of the GARCH models and found that residuals are not heteroscedastic. This means that the time varying structure of the variance is correctly captured by using GARCH models with Gaussian noise. Results of the Engle's test are reported in Table 5, which presents the relative p-values, for each stock and each GARCH model.

We also tested our model for the ability to reproduce the probability mass function of returns. We used a Komogorov–Smirnov (KS) test of hypothesis to verify whether the distribution of real and synthetic data can be considered identical in a statistical sense. We tested the marginal distribution for all 6 stocks and determined that real and synthetic data can be considered as deriving from the same distribution. The distributions are reported in Fig. 4. We also tested whether the multivariate model was able to reproduce the pairwise correlation (cross-correlation) between stocks and still preserve the autocorrelation functions. The definition of cross-correlation between stocks α and β is:

$$\Sigma_{\alpha,\beta} = \frac{Cov(R_\alpha, R_\beta)}{\sqrt{Var(R_\alpha)Var(R_\beta)}}. \tag{14}$$

We estimated the cross-correlation matrix for each pair of stocks from the real time series and from the synthetic ones. Note that the matrix is symmetric with respect to stocks α and β. We show both matrices in Fig. 5 where it is possible to note that the multivariate model is able to almost completely reproduce the cross-correlation between stocks.

Cross-correlation results can be compared with those obtained by the same authors and shown in D'Amico and Petroni (2014) and D'Amico, Petroni, and Prattico (2014). With respect to that model, where only half of the cross-correlation was captured, significantly improved results were obtained by using copula. Therefore, copulas are used not only to simplify the model, but also to improve its performance with respect to simulating real data.

Furthermore, we used the model to forecast the volatility of each stock by using a one step ahead forecasting procedure. The database is divided into two subsets: the first part is used to find the transition probability matrix (as described in the previous

Table 5
p-value of the Engele's ARCH test for residuals of the given GARCH model. The null hypothesis is that residuals are heteroscedastic. In almost all cases the null hypothesis is rejected by the test with a high significance level.

Stock	GARCH(1,1)	GARCH(1,2)	GARCH(2,1)
F	0.64	0.34	0.08
ISP	0.07	0.35	0.00
MS	0.40	0.27	0.00
PC	0.93	0.91	0.82
TIT	0.97	0.87	0.19
TEN	0.16	0.45	0.53

Table 6
Percentage square root mean error on volatility forecasting of each stock for the presented model WISMC, for a simple persistence model and for a GARCH(1,2) model.

Stock	WISMC (%)	Persistence (%)	GARCH(1,2) (%)
F	26	56	33
ISP	28	63	37
MS	23	59	32
PC	29	48	39
TIT	23	54	35
TEN	22	54	30

section), named *setting period*; the second part is used to compare the model forecasting with real data, called *testing period*.

We compare our model with a simple persistence model and a GARCH(1,2) model. The persistence method is often used for its simplicity and for its efficiency for very short-term predictions. It assumes that the volatility at time $t + \Delta t$ is equal to the volatility at time t. Commonly this method is used to compare the behavior of new forecasting models, see e.g. D'Amico et al. (2014). In Table 6 we show the MPE between the real time series and a predicted one by means of the three models. The MPE shown in the table is averaged between the MPE computed on various series obtained as follows: first, we fix the setting period length at 6 months and the time horizon at 100 steps, second we use a sliding window of length 6 months and with 3 months overlap between subsequent series to obtain 20 different sub-series, last we use each time series to set parameters and forecast future values. We ended up with 20 forecasts, each of 100 time horizons that are used to estimate the averaged MPE. It is possible to note from Table 6 that the WISMC model performs better than all other models.

3.3. Results on portfolio risk measures

The multivariate model presented here is used to compare results on some risk measures on a portfolio of 6 stocks. The portfolio is simply built by an equally weighted sum of the 6 analyzed stocks. One of the most important activities in the financial, as well as in the actuarial worlds is assessing the risk of uncertain aggregated positions. This risk is often measured by the Value-at-Risk (VaR_α) at probability level α. VaR_α is the lower α-quantile of the net risk position Y and it is defined as follows:

$$Var_\alpha = \inf\{t \in R; Pr[Y \le t] \ge \alpha\} \tag{15}$$

In this definition of Value-at-Risk, we take the convention of counting losses as positive.

We estimated VaR_α for both real and synthetic data finding almost identical values, results are shown in Fig. 6. The marginal contribution to the VaR_α of each stock has been also estimated and compared in Fig. 7. Also in this case the model is able to exactly reproduce the behavior of real data.

Finally, we compared the Conditional Value-at-Risk (CVaR), also known as expected shortfall, for a level α. The formal definition is the following:

$$CVar_\alpha = \frac{1}{1-\alpha}\int_\alpha^\infty VaR_\gamma \, d\gamma. \tag{16}$$

Comparison of results are shown in Fig. 8.

4. Conclusion

In this paper we proposed a novel multivariate model for high-frequency financial data based on weighted-indexed semi-Markov chains and copulas functions. The model is applied to real data from 6 stocks on the Italian Stock Exchange. Results reveal that the multivariate model is able to reproduce statistical regularities on real data, including the shape of probability mass function of returns, autocorrelation functions and portfolio risk measures. The model was also successfully applied to forecast volatility.

The model relies only on asset returns that are observable variables and is very different from those of the ARCH/GARCH family where the volatility is directly modeled as a correlated process and price is considered to be a collateral effect of the volatility process to which a noise process transformation is applied. Nevertheless, by directly modeling asset returns, the model is able to capture volatility correlation.

Acknowledgment

The authors thank two anonymous referees for their careful revision of the paper and their valuable comments that greatly improved the substance and the presentation of the paper.

Appendix A. Table of notations

Table 7 shows the notations used for the formulation of the mathematical model.

Appendix B. Nonparametric estimation of the weighted-indexed semi-Markov kernel

In this appendix we explain how to derive an empirical estimator of the weighted-indexed semi-Markov kernel. The methodology is based on an extension of the approach advanced in Barbu and Limnios (2006) for ordinary semi-Markov chains.

Let us consider a fixed value of $\lambda \in \mathbb{R}$. Having observed the sequence of discrete returns $\{J_n\}$ and of corresponding jump times $\{T_n\}$ we can calculate through formula (3) the sequence of value of the index process, i.e. $\{V_n^\lambda\}$. This provides a trajectory of the three-dimensional process (J_n, T_n, V_n^λ) observed up to time $L \in \mathbb{N}$:

$$Traj(L) := (J_0, T_0, V_0^\lambda, J_1, T_1, V_1^\lambda, \ldots, J_{N(L)}, T_{N(L)}, V_{N(L)}^\lambda, L - T_{N(L)}, V^\lambda(L)),$$

where $L - T_{N(L)}$ is the censored sojourn time in the last visited state $J_{N(L)}$.

Let $\mathcal{V}_f(L) := \{v \in \mathbb{R} : V_n^\lambda = v, n = 0, 1, \ldots, N(L)\}$ be the set of values assumed by the index process in the observed trajectory $Traj(L)$.

Table 7
List of symbols.

E	Set of discrete returns
Δ	grid amplitude of E
$-z_{\min}\Delta$	minimum discrete return
$z_{\max}\Delta$	maximum discrete return
$S(t)$	Asset price at time t
$R(t)$	log-return at time t
$R_d(t)$	discrete return at time t
$\{J_n\}$	sequence of changing values of discrete returns
$\{T_n\}$	sequence of jump times
$\{V_n^\lambda\}$	sequence of values of the indez process
$Q^\lambda(v;t)$	indexed semi-Markov kernel
$N(t)$	number of transitions by time t
$Z(t)$	indexed semi-Markov chain
$V^\lambda(t)$	index process at time t
$Z_i^{\lambda_{i(t)}}$	return of stock i at time t
$B(t)$	backward recurrence time process at time t
$\mathcal{D}$	set of durations
D	maximum duration in the states
$p_{(i,u,v)}(j,d)$	one-step probability for the WISMC model
$P_{(i,u,v)}(j,d)$	cumulative distribution of asset return
$\mathcal{A}_{(i_h,u_h,v_h)}$	set of admissible states
λ	weight index
s_h	successive adjacent admissible state
$X_h^\lambda(n)$	continuous return index of stock h at time n
$P^{-1}_{(i_h,u_h,v_h)}(\cdot,\cdot)$	quantile function
$R_{(s(n),u(n),v(n))}(x,d)$	joint probability function of the portfolio
$\mathcal{C}$	copula function
$\Sigma(\tau)$	autocorrelation of square returns
$\Sigma_{\alpha,\beta}$	cross-correlation between stock α and β
Var_α	value-at-risk at level α
$CVar_\alpha$	conditional value-at-risk at level α
$Traj(L)$	trajectory of time lenght L of the WISMC model
$\mathcal{V}_f(L)$	set of values of the index process along the trajectory
$N_{ij}(v;t;L)$	counting process of transitions given index value v and sojourn time t
$N_{ij}(v;L)$	counting process of transitions given index value v
$N_i(v;L)$	counting process of visits to state i given index value v
$p_{ij}^\lambda(v)$	indexed transition matrix
$H_i^\lambda(v;t)$	sojourn time distribution function in state i given index value v
$G_{ij}^\lambda(v;t)$	conditional waiting time distribution in the states
$g_{ij}^\lambda(v;t)$	probability mass function of the conditional waiting time distribution
$q_{ij}^\lambda(v;t)$	probability mass function of the weighted-indexed semi-Markov kernel
$\hat{p}_{ij}^\lambda(v;L)$	estimator of the indexed transition matrix
$\hat{g}_{ij}^\lambda(v;t)$	estimator of the probability mass function of the conditional waiting time distribution
$\hat{q}_{ij}^\lambda(v;t)$	estimator of the probability mass function of the weighted-indexed semi-Markov kernel
$Lik(L)$	likelihood function associated to $Traj(L)$
$\mathcal{L}(T)$	Lagrangian function
$\delta_{ij}^{v_h}$	Lagrangian multipliers
$\theta_i^{v_h}$	Lagrangian multipliers

Observe that the set $\mathcal{V}_f(L)$ is finite because the trajectory has a fixed length and that the elements of the set depend on the specific function f considered in formula (3).

Define $\forall i, j \in E$, $t \in \mathbb{N}$ and $v \in \mathcal{V}_f(L)$ the following counting process:

$$N_{ij}(v;t;L) = \sum_{n=1}^{L} 1_{\{J_{n-1}=i, J_n=j, U_{n-1}^m=v, T_n-T_{n-1}=t, T_n\le L\}}. \tag{17}$$

It counts the number of transitions from state i, with an index value v, to state j with a sojourn time in state i equal to t over the interval $[0, L]$.

It is useful also to introduce the processes

$$N_{ij}(v;L) := \sum_{s=1}^{L} N_{ij}(v;s;L); \quad N_i(v;L) := \sum_{j\in E}\sum_{s=1}^{L} N_{ij}(v;s;L) = \sum_{j\in E} N_{ij}(v;L). \tag{18}$$

They calculate the number of transitions from state i with an index value v to state j, up to time L and the number of visits to state i with an index value equal to v, up to time L, respectively.

To derive empirical estimators of the weighted-indexed semi-Markov kernel it is necessary to introduce some auxiliary variables. Let

$$p_{ij}^\lambda(v) := \mathbb{P}[J_{n+1} = j | J_n = i, V_n^\lambda = v],$$

be the transition probabilities of the embedded indexed Markov chain. It denotes the probability that the next transition is in state j given that at current time the process entered in state i and the index process is equal to v. It is simple to realize that $p_{ij}^\lambda(v) = \lim_{t\to\infty} Q_{ij}^\lambda(v;t)$.

Let $H_i^\lambda(v;\cdot)$ be the sojourn time cumulative distribution in state $i \in E$:

$$H_i^\lambda(v;t) := \mathbb{P}[T_{n+1} - T_n \le t | J_n = i,\ V_n^\lambda = v] = \sum_{j\in E} Q_{ij}^\lambda(v;t). \tag{19}$$

It expresses the probability to make a transition from state i with sojourn time less or equal to t given that the indexed process is v.

Define also the conditional waiting time distribution function G:

$$G_{ij}^\lambda(v;t) := \mathbb{P}[T_{n+1} - T_n \le t \mid J_n = i, J_{n+1} = j, V_n^\lambda = v]. \tag{20}$$

It is simple to establish that

$$G_{ij}^\lambda(v;t) = \begin{cases} \dfrac{Q_{ij}^\lambda(v;t)}{p_{ij}^\lambda(v)} & \text{if } p_{ij}^\lambda(v) \neq 0 \\ 1 & \text{if } p_{ij}^\lambda(v) = 0. \end{cases} \tag{21}$$

Now we can define empirical estimators of the probabilities $p_{ij}^\lambda(v)$, of the sojourn time distribution $g_{ij}^\lambda(v;t) := G_{ij}^\lambda(v;t) - G_{ij}^\lambda(v;t-1)$ and of the weighted-indexed semi-Markov kernel $q_{ij}^\lambda(v;t) := Q_{ij}^\lambda(v;t) - Q_{ij}^\lambda(v;t-1)$. They are defined by means of the counting processes defined above, as follows:

$$\hat{p}_{ij}^\lambda(v;L) := \frac{N_{ij}(v;L)}{N_i(v;L)}, \text{ if } \mathrm{N_i(v;L)} \neq 0, \tag{22}$$

$$\hat{g}_{ij}^\lambda(v;t;L) := \frac{N_{ij}(v;t;L)}{N_{ij}(v;L)}, \tag{23}$$

$$\hat{q}_{ij}^\lambda(v;t;L) := \hat{p}_{ij}^\lambda(v;L)\cdot\hat{g}_{ij}^\lambda(v;t;L) = \frac{N_{ij}(v;t;L)}{N_i(v;L)}. \tag{24}$$

In the following part of this appendix we show that the proposed empirical estimators (22)–(24) are approximated maximum likelihood estimators. The likelihood function associated to the trajectory $Traj(L)$ is given by:

$$Lik(L) = \prod_{k=1}^{N(L)} p_{J_{k-1}J_k}^\lambda(V_{k-1}^\lambda) g_{J_{k-1}J_k}^\lambda(V_{k-1}^\lambda; T_k - T_{k-1}) \times\Big(1 - H_{J_{N(L)}}^\lambda(V_{N(L)}^\lambda, L - T_{N(L)})\Big). \tag{25}$$

The quantity $1 - H_{J_{N(L)}}^\lambda(V_{N(L)}^\lambda, L - T_{N(L)})$ can be neglected in order to have an approximated likelihood function:

$$\tilde{Lik}(L) = \prod_{k=1}^{N(L)} p_{J_{k-1}J_k}^\lambda(V_{k-1}^\lambda) g_{J_{k-1}J_k}^\lambda(V_{k-1}^\lambda; T_k - T_{k-1}). \tag{26}$$

This approximation simplifies the maximization of the likelihood function and does not introduce bias because we work with a single trajectory.

The approximated likelihood function can be written using the counting processes (17) and (18) as follows:

$$\begin{aligned}\tilde{Lik}(L) &= \prod_{i,j\in E}\prod_{k\geq 1}\prod_{v_h\in\mathcal{V}_f}\left(p_{ij}^{\lambda}(v_h)g_{ij}^{\lambda}(v_h;k)\right)^{N_{ij}(v_h;k)}\\ &= \prod_{i,j\in E}\prod_{k\geq 1}\prod_{v_h\in\mathcal{V}_f}\left(p_{ij}^{\lambda}(v_h)\right)^{N_{ij}(v_h)}\left(g_{ij}^{\lambda}(v_h;k)\right)^{N_{ij}(v_h;k)}.\end{aligned} \tag{27}$$

Let us maximize the logarithm of the approximated likelihood function:

$$\begin{aligned}log(\tilde{Lik}(L)) = \sum_{i,j\in E}\sum_{k\geq 1}\sum_{v_h\in\mathcal{V}_f}\Big(&N_{ij}(v_h)log(p_{ij}^{\lambda}(v_h))\\ &+N_{ij}(v_h;k)log(g_{ij}^{\lambda}(v_h;k))\Big).\end{aligned} \tag{28}$$

The maximization should be performed under some constraints. The first set of constraints derives from the fact that the rows of the transition probability matrix of the indexed Markov chain sum to 1, i.e. $\sum_{j\in E} p_{ij}^{\lambda}(v_h) = 1$. The second set of constraints derives from the fact that the waiting time in a state given next state are modeled by probability distributions and then the summation of the probability masses should equal 1, i.e. $\sum_{k\geq 1} g_{ij}^{\lambda}(v_h;k) = 1$.

The optimization problem is solved by maximizing the Lagrangian function:

$$\begin{aligned}\mathcal{L}(T) = \sum_{i,j\in E}\sum_{k\geq 1}\sum_{v_h\in\mathcal{V}_f}\Bigg(&N_{ij}(v_h)log(p_{ij}^{\lambda}(v_h)) + N_{ij}(v_h;k)log(g_{ij}^{\lambda}(v_h;k))\\ &+\delta_{ij}^{v_h}\left(1-\sum_{k\geq 1}g_{ij}^{\lambda}(v_h;k)\right)+\theta_i^{v_h}\left(1-\sum_{j\in E}p_{ij}^{\lambda}(v_h)\right)\Bigg)\end{aligned} \tag{29}$$

where the variables $\delta_{ij}^{v_h}$ and $\theta_i^{v_h}$ are the Lagrange multipliers.

By setting the first order partial derivative with respect to $g_{ij}^{\lambda}(v_h;k)$ equal to zero we obtain the condition

$$\frac{N_{ij}(v_h;k)}{g_{ij}^{\lambda}(v_h;k)} - \delta_{ij}^{v_h} = 0 \tag{30}$$

from which we recover $g_{ij}^{\lambda}(v_h;k) = \frac{N_{ij}(v_h;k)}{\delta_{ij}^{v_h}}$. The constraint $\sum_{k\geq 1} g_{ij}^{\lambda}(v_h;k) = 1$ can be rewritten as follows:

$$1 = \sum_{k\geq 1} g_{ij}^{\lambda}(v_h;k) = \sum_{k\geq 1}\frac{N_{ij}(v_h;k)}{\delta_{ij}^{v_h}} = \frac{1}{\delta_{ij}^{v_h}}\sum_{k\geq 1}N_{ij}(v_h;k) = \frac{N_{ij}(v_h)}{\delta_{ij}^{v_h}} \tag{31}$$

from which we obtain $\delta_{ij}^{v_h} = N_{ij}(v_h)$. A simple substitution in (30) gives the empirical estimator $\hat{g}_{ij}^{\lambda}(v;t;L) := \frac{N_{ij}(v;t;L)}{N_{ij}(v;L)}$.

If we calculate the first order derivative of the Lagrangian function with respect to $p_{ij}^{\lambda}(v_h)$ and we make use of the constraint $\sum_{j\in E} p_{ij}^{\lambda}(v_h) = 1$, by means of similar calculations, we obtain the empirical estimator $\hat{p}_{ij}^{\lambda}(v;L) := \frac{N_{ij}(v;L)}{N_i(v;L)}$.

In this way we showed that the empirical estimators (22)–(24) are the approximated nonparametric maximum likelihood estimators.

Appendix C. Algorithm for parameter estimation and optimization

In this appendix, we describe the whole procedure to set and optimize the parameters used in the model.

1. As a first step, to set the WISMC model one has to describe the statistics of the dataset. In particular, the empirical histograms of the distribution can be used to identify $-z_{\min}$ and $z_{\max}$. This procedure is not fully automatized, in our application we decided to use a range such as 99.5% of the data were inside that range;
2. Fix a number of states s to discretize the return variable, fix a value for the weight parameter λ;
3. build the trajectory $(J_n, T_n, V_n^{\lambda})$;
4. Estimate the weighted-indexed semi-Markov kernel $Q_{ij}^{\lambda}(v;t)$ using the empirical estimators derived in Appendix 1;
5. Perform Monte Carlo simulation to build synthetic time series;
6. Estimate the autocorrelation function (ACF) for the synthetic time series $\Sigma(\tau; s, \lambda)$. Note that this ACF depends on the number of states and on the value of the weight parameter;
7. Compare the real ACF, $\Sigma(\tau)$, with the synthetic one, $\Sigma(\tau; s, \lambda)$, by computing the Mean Percentage Error (MPE) between them. The MPE depends on the number of states and on the value of the weight parameter, then it is denoted by MPE(s, λ);
8. Return to point 2, change the number of states and the parameter λ and repeat all points;

At the end of the whole process, choose the number of states s^* and parameter λ^* that best represent the dataset by minimizing the MPE(s, λ), i.e.

$$(s^*, \lambda^*) = \underset{(s,\lambda)}{\text{argmin}}\{\text{MPE}(s, \lambda)\}.$$

Notice that the algorithm can stop whenever the increase in the number of states does not decrease the MPE more than a fixed quantity ϵ.

This procedure should be repeated for all the stocks in the portfolio. Once all the parameters for the univariate models are optimized use a copula to build the multivariate model.

References

Barbu, V. S., & Limnios, N. (2006). Nonparametric estimation for discrete time semi–markov processes with applications in reliability. *Journal of Nonparametric Statistics, 18*(7–8), 483–498.

Bauwens, L., & Hautsch, N. (2009). Modelling financial high frequency data using point processes. In *Handbook of financial time series* (pp. 953–979). Springer Berlin Heidelberg.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics, 141*(2), 876–912.

van Boxtel, E. G. (2007). Modelling credit rating dynamics conditional on counterparty default. Master Thesis, Tilburg University.

Boyle, P. (1986). Option valuation using a three-jump process. *International Options Journal, 3*, 7–12.

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

Chan, J. C. C., & Kroese, D. P. (2010). Efficient estimation of large portfolio loss probabilities in t-copula models. *European Journal of Operational Research, 205*, 361–367.

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). Copula methods in finance. *Wiley finance series*. Wiley, Chichester.

Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics, 7*, 229–263.

Dacorogna, M. M., Genay, R., Muller, U. A., Olsen, R. B., & Pictet, O. V. (2001). *An introduction to high-frequency finance*. San Diego: Academic Press.

D'Amico, G., & Petroni, F. (2011). A semi-Markov model with memory for price changes. *Journal of Statistical Mechanics: Theory and Experiment*, P12009.

D'Amico, G., & Petroni, F. (2012). A semi-Markov model for price returns. *Physica A, 391*, 4867–4876.

D'Amico, G., & Petroni, F. (2012). Weighted-indexed semi-Markov models for modeling financial returns. *Journal of Statistical Mechanics: Theory and Experiment*, P07015.

D'Amico, G., & Petroni, F. (2014). Multivariate high-frequency financial data via semi-Markov processes. *Markov Processes and Related Fields, 20*, 415–434.
D'Amico, G., Petroni, F., & Prattico, F. (2014). Wind speed forecasting at different time scales: a non parametric approach. *Physica A: Statistical Mechanics and its Applications, 406*, 59–66.
Durante, F., Fernández-Sánchez, J., Quesada-Molina, J. J., & Úbeda-Flores, M. (2015). Convergence results for patchwork copulas. *European Journal of Operational Research, 247*(2), 525–531.
Durante, F., & Jaworski, P. (2010). Spatial contagion between financial markets: a copula-based approach. *Applied Stochastic Models in Business and Industry, 26*, 551–564.
Embrechts, P., Hoing, A., & Juri, A. (2003). Using copulae to bound the value-at-risk for functions of dependent risks. *Finance and Stochastics, 7*(2), 145–167.
Engle, R. (1988). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica, 96*, 893–920.
Fodra, P., & Pham, H. (2015). Semi-markov model for market microstructure. *Applied Mathematical Finance, 22*(3), 261–295.
Gorenflo, R., Mainardi, F., Scalas, E., & Raberto, M. (2001). Fractional calculus and continuous-time finance III: the diffusion limit. In M. Kohlmann, & S. Tang (Eds.), *Mathematical finance* (pp. 171–180). Birkhauser, Basel.
Hautsch, N. (2004). *Modelling irregularly spaced financial data*. Springer, Berlin.
Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika, 58*, 83–90.
Joe, H. (1997). *Multivariate dependence concepts*. London: Chapman and Hall.
Kakouris, I., & Rustem, B. (2014). Robust portfolio optimization with copulas. *European Journal of Operational Research, 235*(1), 28–37.
Laih, Y. W. (2014). Measuring rank correlation coefficients between financial time series: A GARCH-copula based sequence alignment algorithm. *European Journal of Operational Research, 232*, 375–382.
Lwin, K. T., Qu, R., & MacCarthy, B. L. (2017). Mean-var portfolio optimization: A nonparametric approach. *European Journal of Operational Research, 260*, 751–766.
Mainardi, F., Raberto, R., Gorenflo, R., & Scalas, E. (2000). Fractional calculus and continuous-time finance II: the waiting-time distribution. *Physica A, 287*, 468–481.
Mansini, R., Ogryczak, W., & Speranza, M. G. (2007). Conditional value at risk and related linear programming models for portfolio optimization. *Annals of Operations Research, 152*(1), 227–256.
Metzler, R., & Klafter, J. (2000). The random walks guide to anomalous diffusion: A fractional dynamics approach. *Physical Report, 339*, 1–77.
Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer.
Sak, H., Hormann, W., & Leydold, J. (2010). Efficient risk simulations for linear asset portfolios in the t-copula model. *European Journal of Operational Research, 202*, 802–809.
Scalas, E., Gorenflo, R., & Mainardi, F. (2000). Fractional calculus and continuous–time finance. *Physica A, 284*, 376–384.
Vassiliou, P. C. G., & Papadopoulou, A. A. (1992). Non-homogeneous semi-Markov systems and maintainability of the state sizes. *Journal of Applied Probability, 29*(3), 376–384.