Stat Papers (2015) 56:1163–1173
DOI 10.1007/s00362-014-0630-6

REGULAR ARTICLE

Estimation and inference in multivariate Markov chains

João Nicolau · Flavio Ivo Riedlinger

Received: 24 October 2012 / Revised: 13 August 2014 / Published online: 2 September 2014

Abstract The literature of Markov chains has recently focused on modeling multiple categorical data sequences. The usual procedure for handling these multivariate Markov chains (MMC), with m categorical data and s states, consists of expanding the state space by considering m^s new states. This model rapidly becomes intractable even with moderate values of m and s due to the excessive number of parameters to estimate. Ching and Fung (2002) found a way to cope with the intractability of the conventional MMC. They also suggested a method of estimation that proved to be inefficient. Zhu and Ching (2010) proposed another method of estimation based on minimizing the prediction error with equality and inequality restrictions. However, both these procedures treat the estimation problem as a mechanic method, without addressing the statistical inference problem. In this article we try to overcome this shortcoming and, at the same time, we propose a new approach to estimate MMC (under Ching et al. hypothesis) which avoids imposing equality and inequality restrictions on the parameters. We illustrate the model and the estimation method with two applications on financial time series data.

Keywords Multivariate Markov chains · Nonlinear least squares · Predictability of investment recommendations · Statistical inference

Mathematics Subject Classification 62M02 · 62M05 · 62M10

J. Nicolau (✉) · F. I. Riedlinger
School of Economics and Management (ISEG), Universidade de Lisboa and CEMAPRE ISEG, Rua do Quelhas 6, 1200-781 Lisbon, Portugal
e-mail: nicolau@iseg.utl.pt

1 Introduction

Markov chains are applied in a number of fields such as physics, chemistry, information sciences, queueing theory, internet, economics and finance, social sciences, biology, etc [more recent applications can be found, for example, in Tsai and Yen (2011) and Faraz and Saniga (2011)]. Recently the literature has focused on modeling multiple categorical data sequences. When the number of categorical data, say s, and the number of states each data can take on, say m, are low, one can expand the state space by considering a first-order Markov chain with m^s states. However, this model rapidly becomes intractable even with moderate values of m and s due to the excessive number of parameters to estimate.

In this context, Ching and Fung (2002) found a way to cope with the intractability of the conventional multivariate Markov chain (MMC) by developing a model with far fewer parameters based on a mixture transition distribution model. This hypothesis was already considered by Raftery (1985) for modeling high-order Markov chains, as an extension of Pegram (1980). This model allows both the intra and inter-transition probabilities among the categorical data. They also propose a method to estimate the parameters based on linear programming. The MMC model has been applied to Markov chain Monte Carlo (Kosorok 2000), demand predictions (Ching and Fung 2002), credit risk (Kijima et al. 2002), reproductive biology (McDonnell et al. 2002), stock markets Maskawa (2003), DNA sequences and Genetic Networks (Ching and Ng 2006), weather simulation (Yang et al. 2011), credit rating (Siu et al. 2002; Fung and Siu 2012).

Recently Zhu and Ching (2010) have proposed a method of estimation based on minimizing the prediction error involving equality and inequality restrictions. They do not address the statistical inference problem (Ching and Fung (2002), do not focus this issue either). Our article has two goals: first we propose a new approach to estimate MMC which avoids imposing equality and inequality restrictions on the parameters, which facilitate the model estimation and the statistical inference. Furthermore, we address the statistical inference of MMC models as proposed by Ching and Fung (2002). We illustrate the model and the estimation method with two applications on financial time series data.

This article is organized as follows: in the next section we present the main results concerning the estimation and inference of MMC. In the last section we illustrate the model and the estimation method with two applications on financial time series data.

2 Estimation statistical inference of multivariate Markov chain

Consider the multivariate stochastic process $\{(S_{1t}, ..., S_{st}) ; t = 1, 2, ...\}$ where S_{jt} $(j = 1, ..., s)$ can take on values on the set $\{1, 2, ..., m\}$. We may rewritten this process as $\left\{\left(\mathbf{x}_t^{(1)}, ..., \mathbf{x}_t^{(s)}\right) ; t = 1, 2, ...\right\}$ where

$$\mathbf{x}_t^{(j)} = \begin{cases} (1,0,0,...,0)' & \text{if } S_{jt} = 1 \\ (0,1,0,...,0)' & \text{if } S_{jt} = 2 \\ \vdots & \vdots \\ (0,0,0,...,1)' & \text{if } S_{jt} = m. \end{cases}$$

Given $S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s$, the kth element of $\mathbf{x}_t^{(j)}$ is a random variable that takes on the value one with probability

$$P\left(S_{jt} = k\big|_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right).$$

Modeling these probabilities using the conventional Markov chain is impracticable since the total number of states of the process increases exponentially with s (there are m^s states). A simplifying hypothesis is considered in Ching and Fung (2002). It involves assuming that the probability $P\left(S_{jt} = k \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right)$ can be written as a convex linear combination of $P\left(S_{jt} = k \middle| S_{1,t-1} = i_1\right), ..., P\left(S_{jt} = k \middle| S_{s,t-1} = i_s\right)$, i.e.

$$\begin{aligned} P\left(S_{jt} = k \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right) = \lambda_{j1} P\left(S_{jt} = k \middle| S_{1,t-1} = i_1\right) \\ + ... + \lambda_{js} P\left(S_{jt} = k \middle| S_{s,t-1} = i_s\right) \end{aligned} \quad (1)$$

where $0 \leq \lambda_{ji} \leq 1$ and $\sum_{i=1}^{s} \lambda_{ji} = 1$. The first approach to estimate the parameters λ_{jk} is described in Ching and Fung (2002). This method solves a minimization problem involving the stationary vector. As referred to in Zhu and Ching (2010), this may imply a large error when the data sequence period is not long enough. This method is certainly not optimal in the mean square error sense (since it does not involve the conditional mean).

We notice that a probability like $P\left(S_{jt} = k \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right)$ is formally identical to the conditional moment

$$\mathrm{E}\left(\mathcal{I}\left(S_{jt} = k\right) \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right),$$

where $\mathcal{I}(A)$ is an indicator function that takes on the value one if A is true. Let $\mathbf{P}^{(jk)}$ be a $m \times m$ matrix with elements $P_{ab}^{(jk)} := P\left(S_{jt} = a \middle| S_{k,t-1} = b\right)$. Therefore

$$\underbrace{\begin{bmatrix} \mathrm{E}\left(\mathcal{I}\left(S_{jt} = 1\right) \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right) \\ \vdots \\ \mathrm{E}\left(\mathcal{I}\left(S_{jt} = m\right) \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right) \end{bmatrix}}_{\mathrm{E}\left(\mathbf{x}_t^{(j)} \middle| S_{1,t-1} = i_1, ..., S_{s,t-1} = i_s\right)}$$

$$= \lambda_{j1} \underbrace{\begin{bmatrix} P_{1i_1}^{(j1)} \\ \vdots \\ P_{mi_1}^{(j1)} \end{bmatrix}}_{i_1\text{th column of } \mathbf{P}^{(j1)}} + \ldots + \lambda_{js} \underbrace{\begin{bmatrix} P_{1i_s}^{(js)} \\ \vdots \\ P_{mi_s}^{(js)} \end{bmatrix}}_{i_s\text{th column of } \mathbf{P}^{(js)}}.$$

Now define $\mathcal{F}_{t-1}$ as all the available information of the system at time $t-1$, i.e. let $\mathcal{F}_{t-1}$ be the σ-algebra generated by $\left\{\left(S_{1,t-1}, \ldots, S_{s,t-1}\right), \left(S_{1,t-2}, \ldots, S_{s,t-2}\right), \ldots\right\}$. This means that the hypothesis (1) can be put in the following terms:

$$\mathrm{E}\left(\mathbf{x}_t^{(j)} \middle| \mathcal{F}_{t-1}\right) = \lambda_{j1}\mathbf{P}^{(j1)}\mathbf{x}_{t-1}^{(1)} + \ldots + \lambda_{js}\mathbf{P}^{(js)}\mathbf{x}_{t-1}^{(s)} \tag{2}$$

where

$$\mathrm{E}\left(\mathbf{x}_t^{(j)} \middle| \mathcal{F}_{t-1}\right) := \begin{bmatrix} \mathrm{E}\left(\mathcal{I}\left(S_{jt}=1\right) \middle| \mathcal{F}_{t-1}\right) \\ \vdots \\ \mathrm{E}\left(\mathcal{I}\left(S_{jt}=m\right) \middle| \mathcal{F}_{t-1}\right) \end{bmatrix}$$

(obviously, $\mathrm{E}\left(\mathcal{I}\left(S_{jt}=k\right) \middle| \mathcal{F}_{t-1}\right) = \mathrm{E}\left(\mathcal{I}\left(S_{jt}=k\right) \middle| S_{1,t-1}, \ldots, S_{s,t-1}\right)$, given the Markovian nature of the process). To illustrate this, assume that $s=3$ (three categorical sequences) and $m=2$. Suppose that at time $t-1$ one observes $S_{1t-1}=1$, $S_{2t-1}=1$ and $S_{3t-1}=2$, i.e.

$$\mathbf{x}_{t-1}^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \mathbf{x}_{t-1}^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \mathbf{x}_{t-1}^{(3)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then the conditional mean of $\mathbf{x}_t^{(1)}$ is

$$\mathrm{E}\left(\mathbf{x}_t^{(1)} \middle| \mathcal{F}_{t-1}\right) = \lambda_{11} \underbrace{\mathbf{P}^{(11)}\mathbf{x}_{t-1}^{(1)}}_{\text{1st col. of } \mathbf{P}^{(11)}} + \lambda_{12} \underbrace{\mathbf{P}^{(12)}\mathbf{x}_{t-1}^{(2)}}_{\text{1st col. of } \mathbf{P}^{(12)}} + \lambda_{13} \underbrace{\mathbf{P}^{(13)}\mathbf{x}_{t-1}^{(3)}}_{\text{2nd col. of } \mathbf{P}^{(13)}}.$$

In practise, the probabilities $\mathbf{P}_{ab}^{(jk)}$ have to be estimated. Consistent estimates can be obtained as follows [see Ching and Fung (2002)]:

$$\hat{P}\left(S_{jt}=a \middle| S_{k,t-1}=b\right) = \frac{\sum_{t=1}^{n}\mathcal{I}\left(S_{jt}=a, S_{k,t-1}=b\right)}{\sum_{t=1}^{n}\mathcal{I}\left(S_{k,t-1}=b\right)}.$$

We now address the estimation of λ_{jk}. An important step in the estimation procedure consists of representing the MMC through an equation with a martingale difference error term, where the standard nonlinear least squares can be readily applied. Using the fact that any process $\mathbf{x}_t^{(j)}$ can be always written as $\mathbf{x}_t^{(j)} = \mathrm{E}\left(\mathbf{x}_t^{(j)} \middle| \mathcal{F}_{t-1}\right) + \mathbf{u}_t^{(j)}$

with $\mathbf{u}_t^{(j)} := \mathbf{x}_t^{(j)} - E\left(\mathbf{x}_t^{(j)} \middle| \mathcal{F}_{t-1}\right)$, we represent the MMC as

$$\mathbf{x}_t^{(j)} = \sum_{k=1}^{s} \lambda_{jk} \mathbf{P}^{(jk)} \mathbf{x}_{t-1}^{(k)} + \mathbf{u}_t^{(j)}, \qquad j = 1, 2, ..., s \tag{3}$$

where $\mathbf{u}_t^{(j)} := \mathbf{x}_t^{(j)} - \mathrm{E}\left(\mathbf{x}_t^{(j)} \middle| \mathcal{F}_{t-1}\right)$ is the prediction error, which by construction is a martingale difference. To impose $\sum_{i=1}^{s} \lambda_{ji} = 1$ we replace λ_{js} by $\lambda_{js} = 1 - \lambda_{j1} - ... - \lambda_{j,s-1}$ in Eq. (3). Rearranging the terms leads to

$$\begin{aligned}\mathbf{x}_t^{(j)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)} = \lambda_{j1} \left(\mathbf{P}^{(j1)} \mathbf{x}_{t-1}^{(1)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)}\right) + \lambda_{j2} \left(\mathbf{P}^{(j2)} \mathbf{x}_{t-1}^{(2)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)}\right) \\ + ... + \lambda_{j,s-1} \left(\mathbf{P}^{(j,s-1)} \mathbf{x}_{t-1}^{(s-1)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)}\right) + \mathbf{u}_t^{(j)}.\end{aligned}$$

To simplify the notation let us rewrite the previous equation as:

$$\begin{aligned}\mathbf{y}_t^{(j)} &= \lambda_{j1} \mathbf{z}_{t-1,1}^{(j)} + \lambda_{j2} \mathbf{z}_{t-1,2}^{(j)} + ... + \lambda_{j,s-1} \mathbf{z}_{t-1,s-1}^{(j)} + \mathbf{u}_t^{(j)} \\ &= \sum_{k=1}^{s-1} \lambda_{jk} \mathbf{z}_{t-1,k}^{(j)} + \mathbf{u}_t^{(j)}\end{aligned}$$

where $\mathbf{y}_t^{(j)} := \mathbf{x}_t^{(j)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)}$ and $\mathbf{z}_{t-1,k}^{(j)} = \left(\mathbf{P}^{(jk)} \mathbf{x}_{t-1}^{(k)} - \mathbf{P}^{(js)} \mathbf{x}_{t-1}^{(s)}\right)$.

We have yet to deal with the restrictions $0 \leq \lambda_{jk} \leq 1$. We consider an approach that consists of replacing the parameters of interest with an auxiliary function that guarantees the validity of the restrictions. Hence, let λ_{jk} be defined as $\lambda_{jk} = h\left(\theta_{jk}\right) = e^{\theta_{jk}} / \left(1 + e^{\theta_{jk}}\right)$, $k = 1, 2, ..., s-1$. By construction $0 \leq \lambda_{jk} \leq 1$, for any value θ_{jk}.

Finally, the model with the restrictions $0 \leq \lambda_{ji} \leq 1$ and $\sum_{i=1}^{s} \lambda_{ji} = 1$ is

$$\mathbf{y}_t^{(j)} = \sum_{k=1}^{s-1} h\left(\theta_{jk}\right) \mathbf{z}_{t-1,k}^{(j)} + \mathbf{u}_t^{(j)}.$$

The idea is simple: we first obtain an estimate for θ_{jk} and then recover the original parameter λ_{jk}, through the equation $\lambda_{jk} = e^{\theta_{jk}} / \left(1 + e^{\theta_{jk}}\right)$. We emphasize that the estimation, as we will see below, is carried out without imposing any kind of restrictions on the parameters θ_{jk}, although the restrictions on the original parameters λ_{jk} are maintained. The estimation of θ_{jk} is clearly a nonlinear least squares (NLS) problem. Define the NLS estimator as

$$\hat{\boldsymbol{\theta}}_j = \arg\min_{\theta_j} \frac{1}{n} \sum_{t=2}^{n} q_t\left(\theta_j\right),$$

where $q_t\left(\theta_j\right) = \sum_{i=1}^{m} \left(y_{it}^{(j)} - \sum_{k=1}^{s-1} h\left(\theta_{jk}\right) z_{i,t-1,k}^{(j)}\right)^2$ and $\theta_j = \left(\theta_{j1}, ..., \theta_{j,s-1}\right)'$. We recall that $\mathbf{y}_t^{(j)}$ is a $m \times 1$ vector. A generic element in this vector is identified as

$y_{it}^{(j)}$, $i = 1, ..., m$. In the same way, a generic element of the $m \times 1$ vector $\mathbf{z}_{t,k}^{(j)}$ is $z_{i,t,k}^{(j)}$. Under some mild regularity conditions, including $\left\{\left(\mathbf{y}_t^{(j)}, \mathbf{z}_{t,k}^{(j)}\right)\right\}$ is a stationary and weakly dependent process, we have

$$\hat{\boldsymbol{\theta}}_j \xrightarrow{p} \theta_j \text{ and } \sqrt{n}\left(\hat{\boldsymbol{\theta}}_j - \theta_j\right) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where $\Sigma = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}$ and

$$\mathbf{A} = \mathrm{E}\left(\frac{\partial^2 q_t(\theta_j)}{\partial\theta_j\theta_j'}\right), \qquad \mathbf{B} = \lim_{n\to\infty}\frac{1}{n}\mathrm{Var}\left(\sum_{t=1}^{n}\frac{\partial q_t(\theta_j)}{\partial\theta_j}\right)$$

(see, for example, Hayashi (2000), Chap. 7).

Remark 1 The assumption that $\left\{\left(\mathbf{y}_t^{(j)}, \mathbf{z}_{t,k}^{(j)}\right)\right\}$ is a stationary and weakly dependent process is a weak condition. Suppose that $P^{(jk)}$ for $j, k = 1, ..., s$ are irreducible and aperiodic. Then $\mathbf{x}_t = \left(\mathbf{x}_t^{(1)}, ..., \mathbf{x}_t^{(s)}\right)'$ has a unique stationary distribution [see Billingsley (1999)]. It follows that a function of stationary process is also a stationary process, thus $\left\{\left(\mathbf{y}_t^{(j)}, \mathbf{z}_{t,k}^{(j)}\right)\right\}$ remains a stationary process. Moreover, all moments of $\mathbf{x}_t$ are bounded in view of the finite state assumption. Due to the fact that all moments are finite, to stationarity and the fact that $\mathbf{x}_t$ is α-mixing with a geometric rate of decay [see Billingsley (1999)], any sample mean follows the law of large numbers and any sum properly standardized has an asymptotic normal distribution by the central limit theorem. These results may be used to easily verify the hypotheses defined in propositions 7.1 and 7.8 of Hayashi (2000) concerning the consistency and asymptotic normality of the NLS estimator.

Remark 2 By the law of large numbers (see previous remark) the matrix $\mathbf{A}$ can be consistently estimated by

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 q_t\left(\hat{\boldsymbol{\theta}}_j\right)}{\partial\theta_j\theta_j'}.$$

To estimate $\mathbf{B}$ we use the fact that

$$\frac{\partial q_t(\theta_j)}{\partial\theta_j} = -\sum_{i=1}^{m}\left(y_{it}^{(j)} - \sum_{k=1}^{s-1} h(\theta_{jk})\, z_{i,t-1,k}^{(j)}\right)\sum_{k=1}^{s-1}\frac{h(\theta_{jk})\, z_{i,t-1,k}^{(j)}}{\partial\theta_j}$$

is a martingale difference (hence its conditional and marginal moment is zero). In effect

$$\mathrm{E}\left(\frac{\partial q_t(\theta_j)}{\partial\theta_j}\middle|\mathcal{F}_{t-1}\right) = -\sum_{i=1}^{m}\mathrm{E}\left(y_{it}^{(j)} - \sum_{k=1}^{s-1} h(\theta_{jk})\, z_{i,t-1,k}^{(j)}\middle|\mathcal{F}_{t-1}\right)\sum_{k=1}^{s-1}\frac{h(\theta_{jk})\, z_{i,t-1,k}^{(j)}}{\partial\theta_j}$$

and $\mathrm{E}\left(y_{it}^{(j)} - \sum_{k=1}^{s-1} h\left(\theta_{jk}\right) z_{i,t-1,k}^{(j)} \middle| \mathcal{F}_{t-1}\right) = 0$ by construction. Moreover, by stationarity we have

$$\frac{1}{n}\mathrm{Var}\left(\sum_{t=1}^{n} \frac{\partial q_t\left(\theta_j\right)}{\partial \theta_j}\right) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\left(\frac{\partial q_t\left(\theta_j\right)}{\partial \theta_j}\frac{\partial q_t\left(\theta_j\right)}{\partial \theta_j'}\right) = \mathrm{E}\left(\frac{\partial q_t\left(\theta_j\right)}{\partial \theta_j}\frac{\partial q_t\left(\theta_j\right)}{\partial \theta_j'}\right).$$

Therefore, a consistent estimator of $\mathbf{B}$ is

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial q_t\left(\hat{\boldsymbol{\theta}}_j\right)}{\partial \theta_j}\frac{\partial q_t\left(\hat{\boldsymbol{\theta}}_j\right)}{\partial \theta_j'}.$$

Given $\hat{\boldsymbol{\theta}}_j$ we recover the parameters of interest:

$$\hat{\lambda}_{jk} = \begin{cases} h\left(\hat{\theta}_{jk}\right) & k = 1, 2, ..., s-1 \\ 1 - \sum_{i=1}^{s-1} h\left(\hat{\theta}_{ji}\right) & k = s \end{cases},$$

Obviously $\hat{\lambda}_{jk} \xrightarrow{p} \lambda_{jk}$ by Slutsky's theorem (given that h is a continuous function). The asymptotic distribution of $\hat{\boldsymbol{\lambda}}_j = \left(\hat{\lambda}_{j1}, ..., \hat{\lambda}_{j,s-1}\right)'$ is given by the delta theorem

$$\sqrt{n}\left(\hat{\boldsymbol{\lambda}}_j - \lambda_j\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{\partial \mathbf{h}\left(\theta_j\right)'}{\partial \theta_j} \Sigma \frac{\partial \mathbf{h}\left(\theta_j\right)}{\partial \theta_j}\right),$$

where $\mathbf{h}\left(\theta_j\right) = \left(h\left(\theta_{j1}\right), ..., h\left(\theta_{j,s-1}\right)\right)'$. In particular, for a scalar estimate $\hat{\lambda}_{ji}$ we have

$$\sqrt{n}\frac{\left(\hat{\lambda}_{ji} - \lambda_{ji}\right)}{\mathrm{Var}\left(\hat{\lambda}_{ji}\right)^{1/2}} \xrightarrow{d} N\left(0, 1\right), \qquad \mathrm{Var}\left(\hat{\lambda}_{ji}\right) = \left(\frac{\partial h\left(\hat{\boldsymbol{\theta}}_{ji}\right)}{\partial \theta_{ji}}\right)^2 \mathrm{Var}\left(\hat{\theta}_{ji}\right).$$

The delta theorem can be applied again to obtain the asymptotic distribution of $\hat{\lambda}_{j,s} = 1 - \sum_{k=1}^{s-1} h\left(\hat{\theta}_{jk}\right)$.

This procedure has to be repeated for other values $j \in \{1, ..., m\}$.

3 Examples

In this section we illustrate the model and the estimation method with two examples from the financial time series area.

MMC is very promising in modeling ratings over time, since ratings have the level of measurement required by MMC models. In the first example we illustrate how MMC may be used to study the predictability of stock or investment recommendations. For

illustration purposes we consider Citigroup's investment recommendation produced by the Bank of America/Merrill Lynch (BofA/ML) research department during the period January 1994-December 2009. Let S_{1t} be the analysts' recommendations, with state space $\{1, 2, 3\}$ defined according to BofA/ML research: $S_{1t} = 1$ (buy) if at time t the annualized return expectation is higher than 10 %; $S_{1t} = 2$ (hold) if at time t the annualized return expectation lays in the interval (0, 10 %) and $S_{1t} = 3$ (sell) otherwise (BofA/ML also considers the "rating dispersion guidelines for coverage cluster"). These recommendations express the particular analyst's opinion about the company's future prospects. To compare the ability of the analysts to correctly predict future returns we also collect Citigroup's monthly returns (since analysts' recommendations are disclosed at the beginning of each month, we consider monthly returns at the beginning of the month too). Monthly returns are split into three categories $\{1, 2, 3\}$ leading to the second stochastic process $\{S_{2t}\}$ to be defined as follows: $S_{2t} = 1$ if at time t the annualized return is higher than 10 %; $S_{2t} = 2$ if at time t the annualized return lays in the interval (0, 10 %) and $S_{2t} = 3$ if otherwise. Let $\hat{\boldsymbol{x}}_t^{(j)}$ be defined as $\hat{\boldsymbol{x}}_t^{(j)} := \sum_{k=1}^{s} \hat{\lambda}_{jk} \hat{\boldsymbol{P}}^{(jk)} \mathbf{x}_{t-1}^{(k)}$ [see Eq. (3)]. Implementing the method described in previous section we obtained:

$$\hat{\boldsymbol{x}}_t^{(1)} = \underset{(0.058)}{0.9907} \hat{\boldsymbol{P}}^{(11)} \mathbf{x}_{t-1}^{(1)} + \underset{(0.058)}{0.009} \hat{\boldsymbol{P}}^{(12)} \mathbf{x}_{t-1}^{(2)} \tag{4}$$

$$\hat{\boldsymbol{x}}_t^{(2)} = \underset{(0.214)}{0.530} \hat{\boldsymbol{P}}^{(21)} \mathbf{x}_{t-1}^{(1)} + \underset{(0.214)}{0.470} \hat{\boldsymbol{P}}^{(22)} \mathbf{x}_{t-1}^{(2)} \tag{5}$$

(standard error in parentheses) where $P_{ab}^{(jk)} := P\left(S_{jt} = a \middle| S_{kt} = b\right)$ and

$$\hat{\boldsymbol{p}}^{(11)} = \begin{bmatrix} 0.86 & 0.06 & 0.09 \\ 0.14 & 0.91 & 0.00 \\ 0.00 & 0.03 & 0.91 \end{bmatrix}, \quad \hat{\boldsymbol{p}}^{(12)} = \begin{bmatrix} 0.35 & 0.35 & 0.27 \\ 0.46 & 0.65 & 0.46 \\ 0.19 & 0 & 0.27 \end{bmatrix}$$

$$\hat{\boldsymbol{p}}^{(21)} = \begin{bmatrix} 0.33 & 0.43 & 0.45 \\ 0.29 & 0.33 & 0 \\ 0.38 & 0.24 & 0.55 \end{bmatrix}, \quad \hat{\boldsymbol{p}}^{(22)} = \begin{bmatrix} 0.42 & 0.35 & 0.41 \\ 0.38 & 0.18 & 0.18 \\ 0.19 & 0.47 & 0.41 \end{bmatrix}.$$

Equation (4) suggests that past returns have little impact on the level of future ratings. Ratings exhibit strong persistence, i.e. BofA/ML research tends to maintain previous ratings (notice that the elements on the diagonal of $\hat{\boldsymbol{p}}^{(11)}$ are relatively high). This is especially true for the ratings $S_{1t} = 2$ (hold) and $S_{t1} = 3$ (sell). On the other hand, Eq. (5) suggests that the analyst stock recommendations may have value for investors (the 0.530 estimate is statistically significant), although the accuracy of those recommendations may not be very high as the following example illustrates. Suppose that in the period $t - 1$ one has $S_{1t-1} = 1$ and $S_{2t-1} = 1$ (there is a recommendation to buy and the annualized return is above 10 %), so $\mathbf{x}_{t-1}^{(1)} = (1, 0, 0)'$ and $\mathbf{x}_{t-1}^{(2)} = (1, 0, 0)'$. Using the estimates of Eq. (5) one obtains $\hat{\boldsymbol{x}}_t^{(2)} = (0.38, 0.33, 0.29)'$. The first entry of $\hat{\boldsymbol{x}}_t^{(2)}$ (value 0.38) represents the probability that the annualized returns are

above 10 % at time t (given $S_{1t-1} = 1$ and $S_{2t-1} = 1$). This probability is relatively low given that there was a recommendation to buy in the previous period. Another interesting scenario is when $S_{1t-1} = 3$ and $S_{2t-1} = 1$ (there was a recommendation to sell and the annualized return was above 10 %), so $\mathbf{x}_{t-1}^{(1)} = (0, 0, 1)'$ and $\mathbf{x}_{t-1}^{(2)} = (1, 0, 0)'$. This scenario involves information with contrarian signs. Using Eq. (5) again we obtain $\hat{\boldsymbol{x}}_t^{(2)} = (0.44, 0.18, 0.38)'$. Despite a sell recommendation, the most likely scenario is a bull market in the following period. This exercise allows us to conclude that an informative recommendation (in the sense that there is a correlation between rating and future returns) may not be accurate enough to present valuable information to investors.

In the second example we consider a multivariate Markov chain to model the SP500, Nikkei 225 and DAX stock indices (we analyze weekly data from January 6, 1965 to March 30, 2011). This example can be seen as a generalization of McQueen and Thorley (1991) approach to analyzing the predictability of stock returns . They consider a Markov chain model to test the random walk hypothesis of stock prices. Their Markov chain is defined by two states: one to represent high returns and the other to represent low returns. Our generalization consists in expanding the number of categorical data (one to three) and the number of states or regimes that each process can take on (we will consider 10 states). The main purpose of our application is only to illustrate how MMC can be used in practice, but several interesting conclusions can be drawn from the data.

Let r_{1t}, r_{2t} and r_{3t} be the returns associated with the SP500, Nikkei 225 and DAX respectively. We split the returns into 10 categories data as follows. Let $q_\alpha^{(i)}$ be the α-quantile of the marginal distribution of r_{it}, i.e. $q_\alpha^{(i)}$ is such that $P\left(r_{it} \leq q_\alpha^{(i)}\right) = \alpha$, and $\hat{q}_\alpha^{(i)}$ the corresponding sample quantile (for simplicity we will refer to the $\hat{q}_{0.10}$ as the 10th percentile, the $\hat{q}_{0.20}$ as the 20th percentile, and so on). We have

$$
\begin{aligned}
S_{it} &= 1 \text{ if } r_{it} \leq \hat{q}_{0.10}^{(i)}, \\
S_{it} &= 2 \text{ if } \hat{q}_{0.10}^{(i)} < r_{it} \leq \hat{q}_{0.20}^{(i)} \\
&\quad \ldots. \\
S_{it} &= 10 \text{ if } r_{it} \geq \hat{q}_{0.90}^{(i)}
\end{aligned}
$$

(the higher the value S_{it} takes on the higher the associated return; for example $S_{1t} = 10$ means that at time t the return of the SP500 index is above the $90th$ percentile). This conversion causes some loss of information. However, Markov chains with more than two states can capture nonlinear dynamics. This is valid for univariate Markov chain models as well as for multivariate model. This feature turns out to be of fundamental importance in our model as we see below. Implementing the method described in the previous section we obtained.

$$
\begin{aligned}
\hat{\boldsymbol{x}}_t^{(1)} &= \underset{(0.072)}{0.351}\,\hat{\boldsymbol{P}}^{(11)}\mathbf{x}_{t-1}^{(1)} + \underset{(0.070)}{0.265}\,\hat{\boldsymbol{P}}^{(12)}\mathbf{x}_{t-1}^{(2)} + \underset{(0.072)}{0.384}\,\hat{\boldsymbol{P}}^{(13)}\mathbf{x}_{t-1}^{(3)} \\
\hat{\boldsymbol{x}}_t^{(2)} &= \underset{(0.075)}{0.182}\,\hat{\boldsymbol{P}}^{(21)}\mathbf{x}_{t-1}^{(1)} + \underset{(0.067)}{0.575}\,\hat{\boldsymbol{P}}^{(22)}\mathbf{x}_{t-1}^{(2)} + \underset{(0.075)}{0.243}\,\hat{\boldsymbol{P}}^{(23)}\mathbf{x}_{t-1}^{(3)}
\end{aligned}
$$

Table 1 Cond. prob. of S_{1t} given $S_{1t-1} = 1$, $S_{2t-1} = 1$, $S_{3t-1} = 1$

1	2	3	4	5	6	7	8	9	10
0.145	0.086	0.103	0.078	0.075	0.067	0.040	0.09	0.102	0.214

Table 2 Cond. prob. of S_{1t} given $S_{1t-1} = 10$, $S_{2t-1} = 10$, $S_{3t-1} = 10$

1	2	3	4	5	6	7	8	9	10
0.123	0.113	0.090	0.092	0.090	0.100	0.084	0.100	0.108	0.100

$$\hat{\boldsymbol{x}}_t^{(3)} = \underset{(0.072)}{0.241}\,\hat{\boldsymbol{P}}^{(31)}\mathbf{x}_{t-1}^{(1)} + \underset{(0.069)}{0.344}\,\hat{\boldsymbol{P}}^{(32)}\mathbf{x}_{t-1}^{(2)} + \underset{(0.072)}{0.415}\,\hat{\boldsymbol{P}}^{(33)}\mathbf{x}_{t-1}^{(3)}$$

(the $\hat{\boldsymbol{P}}^{(ij)}$ matrices of 10×10 dimension, are too large to be presented here. They are available upon request). A striking feature of these results is that all estimates are statistically significant. It is clear from our simulations that the model can anticipate strong increases or decreases in returns. For example, we may ask what the conditional probability function of S_{1t} is given that $S_{1t-1} = 1$, $S_{2t-1} = 1$ and $S_{3t-1} = 1$ (in other words, what are the probabilities associated with S_{1t}, given that all the three returns were below the 10th percentile in the last period). The answer is given by the vector $\hat{\boldsymbol{x}}_t^{(1)}$ (of 10×1 dimension) when all vectors $\mathbf{x}_{t-1}^{(1)}$, $\mathbf{x}_{t-1}^{(2)}$ and $\mathbf{x}_{t-1}^{(3)}$ take on the value 1 at the first entry (and zero in other entries) - see Table 1.

Table 1 shows that the probability of the SP500 being in a bull market (i.e. $S_{1t} = 10$) after the three indices were below the $10th$ percentile in the previous week is relatively high (the probability is 0.214) and higher than the probability of the SP500 continuing below the $10th$ percentile. We may also express this conclusion in terms of the original variables as follows,

$$\begin{aligned} &P\left(r_{1t} > q_{0.90}^{(1)} \middle| r_{1t-1} < q_{0.10}^{(1)}, r_{2t-1} < q_{0.10}^{(2)}, r_{3t-1} < q_{0.10}^{(3)}\right) = 0.214 \\ > &P\left(r_{1t} < q_{0.10}^{(1)} \middle| r_{1t-1} < q_{0.10}^{(1)}, r_{2t-1} < q_{0.10}^{(2)}, r_{3t-1} < q_{0.10}^{(3)}\right) = 0.145. \end{aligned}$$

Another similar exercise can be done, using as conditioning set $S_{1t-1} = 10$, $S_{2t-1} = 10$ and $S_{3t-1} = 10$. The conditional probabilities of S_{1t} are given in Table 2.

Table 2 shows that the probability of the SP500 being in a bear market after the three indices were above the $90th$ percentile in the previous week is relatively higher than the probability of the SP500 continuing above the $90th$ percentile. Both these extreme cases may be related to the famous quotation by Mandelbrot who stated that "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes". Our results go a bit further: not only do they confirm Mandelbrot's idea (that low values of S_{it-1} tend to be followed by low or high

values of values of S_{it}, but not by moderate values) but also enables us to conclude that a bull (bear) market is more likely to be followed by a bear (bull) market.

4 Concluding remarks

We merely illustrate the potential use of the model, but there are several other issues that can be exploited. In fact, since it is quite easy to obtain conditional moments (such as means, variance, skewness and kurtosis) as well as Markov times and marginal moments, many interesting finance applications can be devised in the context of the model. The results also suggest that the model may be able to generate trading rules. This is an issue that may be worth analyzing in a future work.

Acknowledgments This research was supported by the Fundação para a Ciência e a Tecnologia (FEDER/POCI 2010 program)

References

Billingsley P (1999) Convergence of probability measures. John Wiley, New York

Ching W, Fung E (2002) A multivariate Markov chain model for categorical data sequences and its applications in demand predictions. J Manag Math 13:187–199

Ching W, Ng M (2006) Markov chains: models, algorithms and applications, series: international series in operations research & management science, vol 83. Springer, New York

Faraz A, Saniga E (2011) A unification and some corrections to Markov chain approaches to develop variable ratio sampling scheme control charts. Stat Papers 52:799–811

Fung E, Siu T (2012) A flexible Markov chain approach for multivariate credit ratings. Comput Econ 39:135–143

Hayashi F (2000) Econometrics. Princeton University Press, Princeton

Kijima M, Komoribayashi K, Suzuki E (2002) A multivariate Markov model for simulating correlated defaults. J Risk 4:1–32

Kosorok M (2000) Monte Carlo error estimation for multivariate Markov chains. Stat Probab Lett 46:85–93

Maskawa J (2003) Multivariate Markov chain modeling for stock markets. Physica A 324:317–322

McDonnell J, Goverde A, Rutten F, Vermeiden J (2002) Multivariate Markov chain analysis of the probability of pregnancy in infertile couples undergoing. Assist Reprod 17:103–106

McQueen G, Thorley S (1991) Are stock returns predictable? a test using Markov chains. J Financ 46:239–263

Pegram G (1980) An autoregressive model for multilag Markov chains. J Appl Probab 17:350–362

Raftery A (1985) A model for high-order Markov chains. J Royal Stat Soc Series B 47:528–539

Siu T, Ching W, Fung S (2005) On a multivariate Markov chain model for credit risk measurement. Quant Financ 5(6):543–556

Tsai T, Yen W (2011) Exponentially weighted moving average control charts for three-level products. Stat Papers 52:419–429

Yang H, Li Y, Lu L, Qi R (2011) First order multivariate Markov chain model for generating annual weather data for Hong Kong. Energy Build 43:2371–2377

Zhu D, Ching W (2010) A new estimation method for multivariate Markov chain model with application in demand predictions. The 3rd international conference on business intelligence and financial engineering (BIFE), Hong Kong.