Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 41: 1124–1135, 2014
doi: 10.1111/sjos.12087

A New Model for Multivariate Markov Chains

JOÃO NICOLAU

ISEG/ULisboa and CEMAPRE

ABSTRACT. We propose a new model for multivariate Markov chains of order one or higher on the basis of the mixture transition distribution (MTD) model. We call it the MTD-Probit. The proposed model presents two attractive features: it is completely free of constraints, thereby facilitating the estimation procedure, and it is more precise at estimating the transition probabilities of a multivariate or higher-order Markov chain than the standard MTD model.

Key words: high-order Markov chains, maximum likelihood method, mixture transition distribution, multivariate Markov chains

1. Introduction

In this paper, we consider a multivariate stochastic Markov process $\{(S_{1t},\ldots,S_{st});\ t = 1,2,\ldots\}$ where $S_{jt}(j = 1,\ldots,s)$ can take values in the finite set $\{1,2,\ldots,m\}$. One assumes that S_{jt} depends on the previous values of $S_{1t-1},\ldots,S_{jt-1},\ldots,S_{st-1}$, which are used to predict or explain S_{jt}. To simplify the notations, we consider a first-order multivariate Markov chain (MMC), but in the following, S_{jt} can also depend on some explanatory variables lagged over more than one period—our approach may in fact be viewed as a higher-order MMC (we briefly address this issue in Section 4). A natural model to represent dependencies between these categorical variables is the Markov chain, through the transition probabilities $P_j(i_0|i_1,\ldots,i_s) := P(S_{jt} = i_0|S_{1,t-1} = i_1,\ldots,S_{s,t-1} = i_s)$ where $j \in \{1,2,\ldots,s\}$. These probabilities are the main focus of statisticians, and they can be easily estimated through the expression (maximum likelihood estimates)

$$\hat{P}_j(i_0|i_1,\ldots,i_s) = \frac{n_{i_1 i_2 \ldots i_s i_0}}{\sum_{i_0=1}^{n} n_{i_1 i_2 \ldots i_s i_0}}, \tag{1}$$

where $n_{i_1 i_2 \ldots i_s i_0}$ is the number of transitions of type $S_{1,t-1} = i_1,\ldots,S_{s,t-1} = i_s, S_{jt} = i_0$. However, modelling these probabilities, when s and m are relatively large and the sample size is small or even moderate, is impracticable because the total number of parameters is $m^s(m-1)$. In practical terms, this means that the numerator as well as the denominator of (1) may be, in most of cases, zero or very close to zero. As a consequence, the parameters cannot be efficiently estimated or even identified with finite sample size. To overcome this problem, Ching & Fung (2002) considered a simplifying hypothesis, which is, in fact, an extension of Raftery (1985a), for modelling high-order Markov chains (HOMC). It involves assuming that the probability $P_j(i_0|i_1,\ldots,i_s) := P(S_{jt} = i_0|S_{1,t-1} = i_1,\ldots,S_{s,t-1} = i_s)$ can be written as a linear combination of $\{P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)\}$, where $P_{jk}(i_0|i) := P(S_{jt} = i_0|S_{k,t-1} = i)$, that is,

$$\begin{aligned} &P(S_{jt} = i_0|S_{1,t-1} = i_1,\ldots,S_{s,t-1} = i_s) = P_j^{MTD}(i_0|i_1,\ldots,i_s) := \lambda_{j1},\\ &P_{j1}(i_0|i_1) + \ldots + \lambda_{js}P_{js}(i_0|i_s), \end{aligned} \tag{2}$$

where $\sum_{i=1}^{s} \lambda_{ji} = 1$ and

$$0 \leq \sum_{k=1}^{s} \lambda_{jk} P_{jk}(i_0|i_k) \leq 1. \tag{3}$$

This expression is called the mixture transition distribution (MTD) model and tries to combine realism with parsimony (Raftery, 1985a). With $0 \leq \lambda_{ji} \leq 1$, the inequality (3) is automatically satisfied. Imposing this condition has the advantage that the λ parameters may be interpreted as probabilities and that the estimation procedure is easier to implement; however, it reduces the range of dependence patterns, including negative partial effects that the MTD can actually capture.

2. A brief literature review

We first focus on the MTD model and its generalizations, and then on the estimation process.

The MTD model has proven to be very useful in several areas, for example, in wind modelling, social behaviour and DNA sequences, and in many areas of finance and economic areas (see a detailed description of these applications in Berchtold and Raftery, 2002; see also Ching *et al.*, 2004 and Ching *et al.*, 2008). Several generalizations of the HOMC under MTD hypothesis have been proposed aimed at a better data fit and to extend the scope of applications. Raftery (1985b) proposed using different transition matrices for each lag. Berchtold (1996, 1998) generalized this approach. Ching *et al.* (2004) still considered this hypothesis and applied a linear programming formulation to estimate the λ parameters. Mehran (1989a, 1989b) and Le *et al.* (1996) devised an infinite-lag MTD model, which can be useful to capture 'long-memory' effects. Berchtold (1996) discussed a version of an MTD model to analyse missing data. Raftery (1985b) discussed the case of infinite denumerable state spaces. An MTD specification was also generalized to cover the analysis of non-Gaussian processes with an arbitrary state space to model time series exhibiting outliers, change points, bursts of volatility and even flat stretches (see Le *et al.*, 1996). Another extension, considered in Raftery & Banfield (1991), was developed to approximate the conditional distribution of spatial data, in which the temporal reference in the MTD model was replaced by a concept of neighbourhood. Ching *et al.* (2008) combined the HOMC and MMC models in a single model. Other contributions related to the MTD model are made by Adke & Deshmukh (1988), Raftery (1993) and MacDonald & Zucchini (1997), among others.

Let us now focus on the estimation process. To estimate the parameters λ_{ji} of MMC under the MTD hypothesis, Ching & Fung (2002) assumed $0 \leq \lambda_{ji} \leq 1$. They considered a method based on linear programming involving the stationary vector. As referred to in Zhu & Ching (2010), this method generally produces a large error when the data sequence period is not long enough. Zhu & Ching (2010) have proposed a more efficient method based on minimizing the prediction error. However, neither article addresses the statistical inference problem. It is important to emphasize that the maximum likelihood estimation (MLE) for MMC under the MTD hypothesis is essentially the same as the MLE for HOMC under the same hypothesis. In fact, in terms of estimation, the MMC process can be seen as an HOMC if we interpret the conditioning variables $S_{1,t-1}, S_{2,t-1}, \ldots, S_{s,t-1}$ as, respectively, the lagged variables $S_{t-1}, S_{t-2}, \ldots, S_{t-s}$. For this reason, we briefly look at some contributions to the literature on the estimation HOMC under the MTD hypothesis. The log likelihood function is known (either for HOMC or MMC) and is given by

$$\log L = \sum_{i_1 i_2 \ldots i_{i_s} i_0} n_{i_1 i_2 \ldots i_{i_s} i_0} \log\left(P_j^{MTD}(i_0|i_1, \ldots, i_s)\right),$$

subject to $\sum_{i=1}^{s} \lambda_{ji} = 1$ and (3). As referred to by Raftery & Tavaré (1994), this estimation is difficult to carry out as the parameter space is highly non-convex, being defined by a large number of non-linear constraints (in total $m^s(m-1)$). The number of constraints can however be reduced to m. They prove that (3) is equivalent to

$$Tq_-(i) + (1-T)q_+(i) \geq 0 \text{ for all } i, \tag{4}$$

where $T = \sum_{i:\lambda_{ji}\geq 0} \lambda_{ji}$, $q_-(i) = \min_{1\leq g\leq m} P_{jk}(i|g)$ and $q_+(i) = \max_{1\leq g\leq m} P_{jk}(i|g)$.

The maximization of the likelihood, even under the constraints (4), still poses difficulties as the objective function is highly non-linear and the number of constraints can still be considered high. In particular, reaching a global maximum can be difficult, especially if the initial values are far away from the optimal values. Berchtold (2001) proposed a method to improve the selection of the initial values by computing a measure of the strength of the association between each lagged value and the present one. Other papers such as Mehran (1989a) and Berchtold (1998) have also addressed the choice of initial values. Several other strategies have been employed to circumvent the difficulties in maximizing the likelihood given the non-linearity of the objective function and the high number of constraints. Berchtold (2001) developed an algorithm that does not require any 'external optimization routine' and can lead to satisfactory results provided that good initial values are chosen. The idea leads to a modification of the Newton methods and consists of balancing an increase in one of the parameters with an equal decrease in another using the boundary adjustment in the MLE. Lèbre & Bourguignon (2008) also pointed out that '[...] the efficiency for the MTD parameter estimations proposed up to date still remains problematic on account of the large number of constraints on the parameters'. They used the expectation–maximization algorithm to estimate the parameters of the MTD model, with good results, although Chen and Lio mentioned that the complexity from the counts of the pattern of sequences is still unsolved in the search for a global maximizer. Chen & Lio (2009) proposed transforming the non-linear constraints of the parameters in the MTD into box-constraints in that each parameter is given a lower and/or upper bound. This technique allows the MLE to be obtained via a hybrid algorithm from the evolutionary algorithms and/or quasi-Newton algorithms and has the advantage of focusing on a search for a global maximizer.

3. The mixture transition distribution-Probit model

3.1. Motivation

We have shown the usefulness of the MTD and its extensions. One of the main challenges in applying the MTD model is linked to the estimation and the way the non-linear constraints are dealt with in the numerical optimization, although some progress has been made as we described in the previous section (e.g. Berchtold, 2001, Lèbre and Bourguignon, 2008 and Chen and Lio, 2009). However, the constraints associated with the MTD model still pose difficulties. Even in Chen & Lio (2009), who transformed the non-linear constraints of the parameters in the MTD into box-constraints, the constraints are still present.

In this paper, we propose a specification, inspired by the MTD model, which is completely free from constraints, facilitating the estimation procedure and, at the same time, as we show in the succeeding text, is a more accurate specification for $P_j(i_0|i_1,\ldots,i_s)$. We suggest modelling $P_j(i_0|i_1,\ldots,i_s)$ as follows

$$P_j(i_0|i_1,\ldots,i_s) = P_j^{\Phi}(i_0|i_1,\ldots,i_s) := \frac{\Phi(\eta_{j0}+\eta_{j1}P_{j1}(i_0|i_1)+\ldots+\eta_{js}P_{js}(i_0|i_s))}{\sum_{k=1}^{m}\Phi(\eta_{j0}+\eta_{j1}P_{j1}(k|i_1)+\ldots+\eta_{js}P_{js}(k|i_s))}, \quad (5)$$

where $\eta_{ji} \in \mathbb{R}(j = 1,\ldots,s; i = 1,\ldots,m)$ and Φ is the (cumulative) standard normal distribution function. We denote this specification as an MTD-Probit model. We have the following remarks:

(i) the numerator of (5) follows the same principle as the original MTD model: the argument of $\Phi(\cdot)$ is a linear combination of probabilities $P_{jk}(i_0|i_k), k = 1,\ldots,s$, just as in the MTD model;
(ii) no constraints are needed in (5), as $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ is bounded in the interval $(0,1)$, regardless of the values η_{js};
(iii) the purpose of the denominator in (5) is to guarantee that $\sum_{i_0=1}^{m} P_j^{\Phi}(i_0|i_1,\ldots,i_s) = 1$. Notice, by analogy, that the same condition has to hold for $P_j(i_0|i_1,\ldots,i_s)$; that is, $\sum_{i_0=1}^{m} P_j(i_0|i_1,\ldots,i_s) = 1$;
(iv) a constant term η_{j0} is introduced in the $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ specification, and in this way, the proposed specification involves one additional parameter in comparison with the MTD case; although it can be set to zero, η_{j0} generally improves the fit (i.e. allows the probability $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ to be closer to $P_j(i_0|i_1,\ldots,i_s)$);
(v) here Φ can be replaced by another distribution function of any continuous random variable with state space $\mathbb{R}$;
(vi) in principle, it is possible to add exogenous explanatory variables to the model (this topic deserves further research);
(vii) when S_{jt} is the dependent variable, the likelihood is

$$\log L = \sum_{i_1 i_2 \ldots i_{i_s} i_0} n_{i_1 i_2 \ldots i_{i_s} i_0} \log\left(P_j^{\Phi}(i_0|i_1,\ldots,i_s)\right), \quad (6)$$

and the maximum likelihood estimator is defined, as usual, as $\hat{\eta}_j = \arg\max_{\eta_{j1},\ldots,\eta_{js}} \log L$. The parameters $P_{jk}(i_0|i_1), k = 1,\ldots,s$ can be estimated in advance, through the consistent estimators

$$\hat{P}_{jk}(i_0|i_1) = \frac{n_{i_1 i_0}}{\sum_{i_0=1}^{n} n_{i_1 i_0}},$$

where $n_{i_1 i_0}$ is the number of transitions from $S_{k,t-1} = i_1$ to $S_{jt} = i_0$. This procedure greatly simplifies the estimation procedure and does not alter the consistency of the MLE $\hat{\eta}_j$ estimator, as $\hat{P}_{jk}$ is a consistent estimator of P_{jk}.

Equation (5) can be superior to the MTD hypothesis for several reasons. First, in the absence of constraints, the estimation is much easier, and standard numerical optimization routines may apply. We have used the constrained maximum likelihood module in GAUSS software (Aptech Systems, Chandler, Arizona, United States) that allows switching between several algorithms (BFGS, Broyden-Fletcher-Goldfarb-Shanno, DFP, Davidon-Fletcher-Powell, Newton, BHHH, Berndt-Hall-Hall-Hausman, scaled BFGS and scaled DFP) depending on three measures of progress, change in function value, number of iterations or change in line search step length. However, the likelihood (6) is not a strictly concave function on the entire parameter state space; hence, the choice of the starting values is relevant. Second, because no restrictions on the parameters are needed, the MTD-Probit enables the description of a wide range of possible dependencies; according to the theorem in the succeeding text, this range is likely to be wider than that of the MTD. Third, the proposed model is more accurate than the

MTD model in the sense that $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ is closer in Euclidean distance to the true probability $P_j(i_0|i_1,\ldots,i_s)$ than that of $P_j^{MTD}(i_0|i_1,\ldots,i_s)$. This result is proved in the following theorem.

Theorem 1. *Suppose that S_{jt} and $S_{k,t-1}$ are not independent (the transition probability matrices between S_{jt} and $S_{k,t-1}$ do not have identical rows). For each $j \in \{1,\ldots,s\}$, we have*

$$\min_{\eta_{ji}} \sum_{i_1 i_2 \ldots i_s i_0=1}^{m} \left| P_j(i_0|i_1,\ldots,i_s) - P_j^{\Phi}(i_0|i_1,\ldots,i_s) \right|^2 \leq, \tag{7a}$$

$$\min_{\substack{\lambda_{j1}+\ldots+\lambda_{js}=1 \\ 0\leq\sum_{k=1}^{s}\lambda_{jk}P_{jk}(i_0|i_k)\leq 1}} \sum_{i_1 i_2 \ldots i_s i_0=1}^{m} \left| P_j(i_0|i_1,\ldots,i_s) - P_j^{MTD}(i_0|i_1,\ldots,i_s) \right|^2. \tag{7b}$$

Proof. To simplify the notations, consider without any loss of generality that $\eta_i = \eta_{ji}$ and $\lambda_i = \lambda_{ji}$. The probabilities $P_j(i_0|i_1,\ldots,i_s)$ and $P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)$ are assumed to be known for all permutations in the set $\{i_0,i_1,\ldots,i_s\}$. The constraints $0 \leq \sum_{k=1}^{s} \lambda_k P_k(i_0|i_k) \leq 1$ are considered in part (4), below. For now, assume that $\{\lambda_{ji} : \sum_{i=1}^{s} \lambda_{ji} = 1\}$. We prove the theorem in four steps.

(1) The value of the expression of the right-hand side of the inequality (7b) is equal to the sum of squared residuals (SSR) of the regression

$$P_j(i_0|i_1,\ldots,i_s) = \beta_1 P_{j1}(i_0|i_1) + \ldots + \beta_{s-1} P_{j,s-1}(i_0|i_{s-1}) + \beta_s P_{js}(i_0|i_s) + error_1,$$

subject to the restrictions $\sum_{i=1}^{s} \beta_i = 1$. (notes: (i) in classical linear regression terms, $P_j(i_0|i_1,\ldots,i_s)$ may be understood as the ' independent' variable and can take on m^{s+1} values (as many as the number of permutations in the set $\{i_0,i_1,\ldots,i_s\}$). For each of those values, $\{P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)\}$ are the corresponding 'explanatory variables'. (ii) The error term $error_1$ results from the fact that the probabilities $P_j(i_0|i_1,\ldots,i_s)$ are not generally equal to a linear combination of $\{P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)\}$. This linear combination is only an approximation to the true probabilities $P_j(i_0|i_1,\ldots,i_s)$. Hence there is always an error that is identified here by $error_1$). Given that $\beta_s = 1 - \beta_1 - \ldots - \beta_{s-1}$, we may rewrite the previous equation as

$$\begin{aligned} P_j(i_0|i_1,\ldots,i_s) &= P_{js}(i_0|i_s) + \beta_1 P_{j1}(i_0|i_1) + \ldots + \beta_{s-1} P_{j,s-1}(i_0|i_{s-1}) \\ &\quad + (-\beta_1 - \ldots - \beta_{s-1}) P_{js}(i_0|i_s) + error_1, \text{ or} \end{aligned}$$

$$\begin{aligned} P_j(i_0|i_1,\ldots,i_s) - P_{js}(i_0|i_s) &= \beta_1 P_{j1}(i_0|i_1) + \ldots + \beta_{s-1} P_{j,s-1}(i_0|i_{s-1}) \\ &\quad + (-\beta_1 - \ldots - \beta_{s-1}) P_{js}(i_0|i_s) + error_1. \end{aligned} \tag{8}$$

(2) To deal with the left-hand side expression (7a), we use the Gauss–Newton method to find the non-linear regression estimates by running successive linear regressions until a solution is reached. We start by linearizing $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ using a Taylor series expansion with linear terms $P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)$ around the vector $\eta^{(0)}$ such that $\Phi(\eta^{(0)}) = P_{js}(i_0|i_s)$. This produces a linear regression equation of type

$$\begin{aligned} P_j(i_0|i_1,\ldots,i_s) &= P_{js}(i_0|i_s) + \beta_1 P_{j1}(i_0|i_1) + \ldots \\ &\quad + \beta_{s-1} P_{j,s-1}(i_0|i_{s-1}) + \beta_s P_{js}(i_0|i_s) + error_2, \end{aligned}$$

$$P_j(i_0|i_1,\ldots,i_s) - P_j(i_0|i_s) = \beta_1 P_{js}(i_0|i_1) + \ldots + \beta_{s-1} P_{j,s-1}(i_0|i_{s-1}) + \beta_s P_{js}(i_0|i_s) + error_2, \quad (9)$$

where $\beta_1,\ldots,\beta_2$ are unknown parameters, depending on η_i, which are estimated by ordinary least squares. The main point is that the SSR of regression (9) is lower than the SSR of regression (4), despite the fact that both equations use the same ' explanatory variables' $\{P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)\}$. The reason for this difference is that the parameters of (4) are subject to restrictions, whereas the parameters of (9) are free. In other words, a solution of an unconstrained optimization problem is always equal or better than that of a constrained optimization problem. Let $\eta^{(1)}$ be the least squares estimates of (9). The Gauss–Newton algorithm proceeds by approximating $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ through a Taylor series expansion with linear terms $P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)$ around the vector obtained in the previous step, $\eta^{(1)}$, and a new regression is formed.

(3) Now it is necessary to show that successive iterations of the Gauss–Newton method cannot worsen the solution obtained in step (2). A sufficient condition is that (a) the set $\{\eta : F(\eta) \leq F(\eta^{(0)})\}$ is bounded, where $F(\eta) := \sum_{i_1 i_2 \ldots i_{i_s} i_0 = 1}^{m} \left| P_j(i_0|i_1,\ldots,i_s) - P_j^{\Phi}(i_0|i_1,\ldots,i_s;\eta) \right|^2$ and that (b) the Jacobian $J(\eta) := \partial P_j^{\Phi}(i_0|i_1,\ldots,i_s)/\partial\eta$ has full rank in all steps (see, for example, Madsen *et al.*, 2004). Condition (a) may be easily satisfied if one assumes that η is compact (i.e. we assume that any admissible value for η_i is finite). On the other hand, one is able to show that the assumption of the theorem guarantees condition (b) (note: if $S_{k,t-1}$ is independent of S_{jt}, the variable $S_{k,t-1}$ can be removed from the model, and the assumption of the theorem may hold with respect to the other explanatory variables).

(4) The theorem was proven assuming that $\lambda_{j1},\ldots,\lambda_{js}$ belong to the set $\{\lambda_{ji} : \sum_{i=1}^{s} \lambda_{ji} = 1\}$. Therefore, a fortiori, it also applies to the smaller set $\{\lambda_{ji} : \sum_{i=1}^{s} \lambda_{ji} = 1, 0 \leq \sum_{k=1}^{s} \lambda_{jk} P_{jk}(i_0|i_k) \leq 1\}$.

□

The previous theorem does not quantify the gains in using the model $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$. These gains can be small or substantial depending on the values $P_j(i_0|i_1,\ldots,i_s)$ and $\{P_{j1}(i_0|i_1),\ldots,P_{js}(i_0|i_s)\}$. The following example illustrates the gains that can be obtained in using the proposed specification. Consider an MMC $\{(S_{1t}, S_{2t})\}$ with $s = 2$ and $m = 2$. Each process takes values in the set $\{1, 2\}$. Suppose that the data generating process is defined as follows:

$$\begin{aligned}
&P_1(1|1,1) = P(S_{1t} = 1|S_{1,t-1} = 1, S_{2,t-1} = 1) = 0.1, && P_1(2|1,1) = 1 - P_1(1|1,1) = 0.9,\\
&P_1(1|1,2) = P(S_{1t} = 1|S_{1,t-1} = 1, S_{2,t-1} = 2) = 0.1, && P_1(2|1,2) = 1 - P_1(1|1,2) = 0.9,\\
&P_1(1|2,1) = P(S_{1t} = 1|S_{1,t-1} = 2, S_{2,t-1} = 1) = 0.2, && P_1(2|2,1) = 1 - P_1(1|2,1) = 0.8,\\
&P_1(1|2,2) = P(S_{1t} = 1|S_{1,t-1} = 2, S_{2,t-1} = 2) = 0.9, && P_1(2|2,2) = 1 - P_1(1|2,2) = 0.1,
\end{aligned}$$

and $P(S_{r,t-1} = i_2|S_{k,t-1} = i_1) = 0.5$ for $i_2, i_1, k, r \in \{1, 2\}$. By the law of total probability, we obtain the following values for $P_{j1}(i_0|i_1)$ and $P_{j2}(i_0|i_2)$:

$$\begin{aligned}
&P_{11}(1|1) = 0.1,\ P_{11}(2|1) = 0.9,\ P_{11}(1|2) = 0.55,\ P_{11}(2|2) = 0.45,\\
&P_{12}(1|1) = 0.15,\ P_{12}(2|1) = 0.85,\ P_{12}(1|2) = 0.5,\ P_{12}(2|2) = 0.5.
\end{aligned}$$

Given $P_{j1}(i_0|i_1)$ and $P_{j2}(i_0|i_2)$, the precision of $P_1^{\Phi}(i_0|i_1,i_2)$ and $P_1^{MTD}(i_0|i_1,i_2)$ can be compared with the true values $P_1(i_0|i_1,i_2)$, by considering the following optimization problems:

$$\min_{\eta_{1i}} \sum_{i_1 i_2, i_0=1}^{2} \left| P_1(i_0|i_1,i_2) - P_1^{\Phi}(i_0|i_1,i_2) \right|^2 = 0.040;$$

$$\min_{\lambda_{11}+\lambda_{12}=1} \sum_{i_1 i_2=1}^{2} \left| P_1(i_0|i_1,i_2) - P_1^{MTD}(i_0|i_1,i_2) \right|^2 = 0.398.$$

In the second optimization problem, we checked that all estimated values of $P_1^{MTD}(i_0|i_1,i_2)$ were probabilities. There is a significant difference between both methods. Our hypothesis leads to an error that is about 10 times lower than the MTD ethod. This difference obviously depends on the parameters that were previously defined (other values may lead to smaller differences).

3.2. Monte Carlo experiment

We have just performed a numerical analysis to show how close $P_j^{\Phi}(i_0|i_1,\ldots,i_s)$ can be to the true probability. This analysis was conducted after we fixed the values of $P_1(i_0|i_1,i_2)$, $P_1(i_0|i_1)$ and $P_1(i_0|i_1)$ and then deduced the best numerical approximations of $P_1^{MTD}(i_0|i_1,i_2)$ and $P_1^{\Phi}(i_0|i_1,i_2)$ to $P_1(i_0|i_1,i_2)$. It is also interesting to perform a Monte Carlo simulation experiment in which the categorical data are simulated and then the estimates from both methods are compared with the true probabilities. We consider a simple process with two categorical data ($s=2$) and $m=2$ (each variable takes on 1 or 2). Our objective is to estimate $P_1(i_0|i_1,i_2)$ from the maximum likelihood estimates $\hat{P}_1^{\Phi}(i_0|i_1,i_2)$ and $\hat{P}_1^{MTD}(i_0|i_1,i_2)$. Because the results are sensitive to the values of $P_1(i_0|i_1,i_2)$, we let these probabilities take several different values in the set [0,1], as described in the succeeding text. We use the following algorithm:

Step 0: Set $\delta_i = 0.1, i = 1,2,\ldots,6$.
Step 1: Set

$$P_1(1|1,1)=\delta_1,\ P_1(1|1,2)=\delta_2,\ P_1(1|2,1)=\delta_3,\ P_1(1|2,2)=\delta_4,$$
$$p_{11}=\delta_5, \qquad p_{21}=\delta_6$$

(we explain the parameters p_{11} and p_{21} in the succeeding text).
Set 2: Simulate a path $\{(S_{1t},S_{2t})\}, t=1,2,\ldots,n$.

Step 2.1: Initialize the process $\{(S_{1t},S_{2t})\}$.
Step 2.2: Simulate a random variable $u \sim U(0,1)$. Assume that $S_{1,t-1}=i_1$ and $S_{2,t-1}=i_2$. Then $S_{1t}=1$ if $u \le P_1(1|i_1,i_2)$, and $S_{1t}=2$ otherwise.
Step 2.3: Simulate S_{2t} according to the probabilities $P(S_{2t}=i|S_{1t}=j)=p_{ji}$ (say) (note: because we are not focusing on the probability $P_2(i_0|i_1,i_2)$, we simulate S_{2t} from a simple probabilistic structure.
Step 2.4: Return to step 2, until $t=n$.

Step 3: Given the simulated sequence $\{(S_{1t},S_{2t})\}$, estimate the parameters λ_{1i} and η_{1i} by maximum likelihood and obtain , from them, $\hat{P}_1^{MTD}(i_0|i_1,i_2)$ and $\hat{P}_1^{\Phi}(i_0|i_1,i_2)$. If the constraints $0 \le \sum_{k=1}^{s} \hat{\lambda}_{jk}\hat{P}_{jk}(i_0|i_k) \le 1$ are not satisfied, the simulated sequence is removed and not considered in the analysis. (note: in our Monte Carlo study, the aforementioned constraints were satisfied in about 98.5 per cent of cases)

Table 1. *Monte Carlo results*

n	$\frac{\text{Average of}\psi^{MTD}}{\text{Average of}\psi^{\Phi}}$
100	1.10
1000	1.20
5000	1.23

Step 4: Assess the precision of $\hat{P}_1^{MTD}(i_0|i_1,i_2)$ and $\hat{P}_1^{\Phi}(i_0|i_1,i_2)$ by comparing them with the values $P_1(i_0|i_1,i_2)$ defined in step 1, using the statistics

$$\psi^{MTD} = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\left(\hat{P}_1^{MTD}(i_0|i_1,i_2) - P_1(i_0|i_1,i_2)\right)^2,$$

$$\psi^{\Phi} = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\left(\hat{P}_1^{\Phi}(i_0|i_1,i_2) - P_1(i_0|i_1,i_2)\right)^2.$$

Step 5: Increase one δ by 0.1. Keep all others δ_i with the same value. Stop the procedure if $\delta_1 = \ldots = \delta_6 = 0.9$, otherwise go to step 1.

Each parameter takes on nine different values in the range [0.1,0.9]; hence there are $9^6 = 531,441$ permutations. For each of these permutations, we simulate a path $\{(S_{1t}, S_{2t})\}$ with 100, 1000 and 5000 observations (Table 1). To assess the models, we computed a global average of the statistics mentioned in step 4.

Table 1 shows that the differences between the models are not so great as we saw in the numerical analysis. Nevertheless, it is clear that the estimator $\hat{P}^{\Phi}$ dominates the $\hat{P}^{MTD}$.

3.3. An empirical application

In this section, we illustrate our method by considering an MMC to model the SP500, Nikkei 225 and DAX stock indices (we analyse weekly data from 6 January 1965 to 5 December 2012, which corresponds to 2289 observations). This example can be seen as a generalization of McQueen & Thorley (1991) approach to analysing 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. We generalize this approach by considering three categorical data ($s = 3$) and ten states ($m = 10$). A fully parameterized MMC involves $m^s(m-1) = 9000$ independent parameters, which is impossible to estimate with only 2289 observations. The main purpose of this application is only to illustrate the proposed model and to compare both methods.

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 ten categories as follows. Let $q_\alpha^{(i)}$ be the α-quantile of the marginal distribution of r_{it}; that is, $q_\alpha^{(i)}$ is such that $P\left(r_{it} \le q_\alpha^{(i)}\right) = \alpha$, and $\hat{q}_\alpha^{(i)}$ the corresponding sample quantile (for simplicity, we will refer to the $\hat{q}_{0.10}$ as the tenth percentile, the $\hat{q}_{0.20}$ as the 20th percentile and so on). We have

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

(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).

Tables 2 and 3 present the estimation results of both methods described in the previous section (in the MTD case, we ran the optimization procedure with no restrictions on the λ terms. In all cases, the restrictions (3) were satisfied).

These results show that the proposed model is superior to that of the MTD model, both in terms of likelihood and Bayesian information criterion ($BIC = -2LL + q\log(n)$, where LL is the log likelihood, q represents the number of independent parameters and n the sample size), despite the fact that our model has one additional parameter (the data and the routines in GAUSS to estimate the models are available at site: http://pascal.iseg.utl.pt/~nicolau/myHP/codes.rar). An interesting fact is that all estimates are statistically significant. This means that both models may have predictive power.

We present a simple illustration of the famous quotation by Mandelbrot when referring to returns behaviour: 'large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes.' Suppose that in the previous period, all three returns were below the tenth percentile (there is a large negative change in period $t-1$). Then, from expression P_j^{Φ} and estimates $\hat{\eta}_{jk}$, we may calculate the conditional probabilities $\hat{P}_1^{\Phi}(i_0|i_1 = 1, i_2 = 1, i_3 = 1)$ (Table 4).

Table 4 shows that the probability of the SP500 being in a bull market (i.e. $S_{1t} = 10$) after the three indices were below the tenth percentile in the previous week is relatively high (the probability is 0.3124) and higher than the probability of the SP500 continuing below the tenth percentile. Another similar exercise can be performed, using the 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 5.

Table 2. *Results of the mixture transition distribution model*

	$\hat{\lambda}_{j1}$	$\hat{\lambda}_{j2}$	$\hat{\lambda}_{j3}$	log $Lik.$	BIC
Equation 1 (SP500, $j = 1$)	0.2777 (0.0788)	0.3274 (0.0779)	0.3949 (0.0781)	−1178.44	2380.08
Equation 2 (Nikkei 225, $j = 2$)	0.2609 (0.0789)	0.5838 (0.0690)	0.1553 (0.0823)	−1177.48	2378.16
Equation 3 (DAX, $j = 3$)	0.2311 (0.0779)	0.3889 (0.0743)	0.3800 (0.0776)	−1179.90	2383.00

BIC, Bayesian information criterion.

Table 3. *Results of the proposed model*

	$\hat{\eta}_{j0}$	$\hat{\eta}_{j1}$	$\hat{\eta}_{j2}$	$\hat{\eta}_{j3}$	log $Lik.$	BIC
Equation 1 (SP500, $j = 1$)	−2.6524 (0.1623)	6.7873 (1.2826)	7.3376 (1.3102)	7.094 (1.3173)	−1166.78	2364.50
Equation 2 (Nikkei 225, $j = 2$)	−3.4530 (0.6657)	2.6336 (0.8004)	2.5880 (0.7430)	2.5880 (0.7430)	−1165.93	2362.80
Equation 3 (DAX, $j = 3$)	−3.0819 (0.2770)	9.284 (1.7169)	9.8165 (1.7544)	9.3397 (1.724)	−1166.32	2363.58

BIC, Bayesian information criterion.

Table 4. *Estimates* $\hat{P}_1^{\Phi}(i_0|i_1 = 1, i_2 = 1, i_3 = 1)$

1	2	3	4	5	6	7	8	9	10
0.2135	0.078	0.0748	0.0469	0.0306	0.0314	0.011	0.0956	0.1059	0.3124

Table 5. *Estimates* $\hat{P}_1^{\Phi}(i_0|i_1 = 10, i_2 = 10, i_3 = 10)$

1	2	3	4	5	6	7	8	9	10
0.1424	0.1396	0.1038	0.068	0.0899	0.097	0.0927	0.0808	0.1042	0.0814

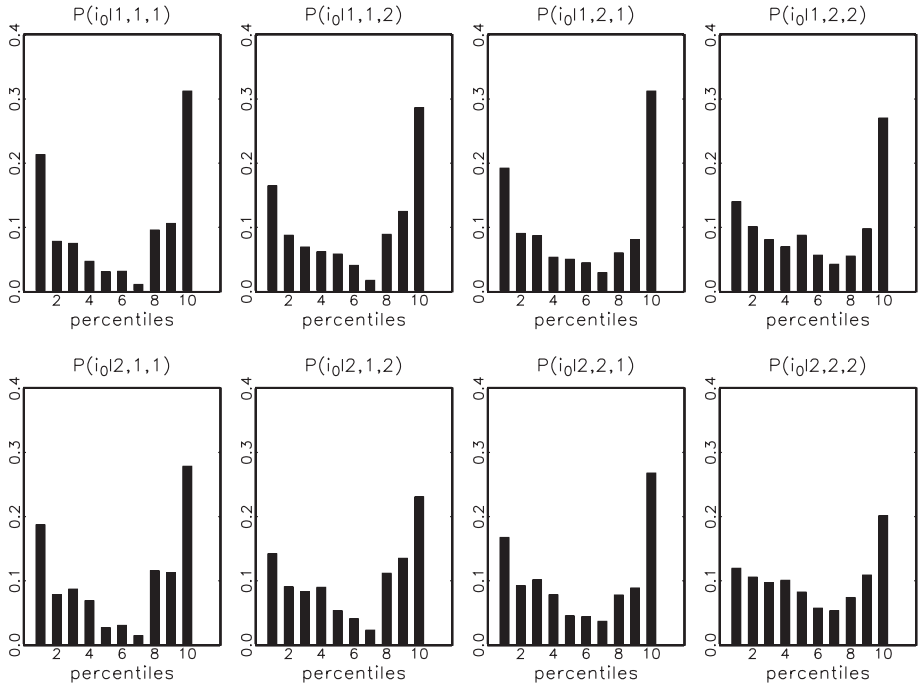

Fig. 1. Conditional probabilities $\hat{P}_1^{\Phi}(i_0|S_{1t-1}, S_{2t-1}, S_{3t-1}); , 2\}$.

Table 5 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 high and higher than the probability of the SP500 continuing above the 90th percentile. Our results not only confirm Mandelbrot's idea (that low values of S_{it-1} tend to be followed by low or high values of S_{it}, but not by moderate values) but also enable us to conclude that a bull (bear) market is more likely to be followed by a bear (bull) market. This conclusion is also confirmed by Fig. 1. In the first panel of this figure, we plot $\hat{P}_1^{\Phi}(i_0|i_1 = 1, i_2 = 1, i_3 = 1)$ (i.e. the values of Table 4). In the second panel, we plot $\hat{P}_1^{\Phi}(i_0|i_1, i_2, i_3)$ when $S_{1t-1}S_{2t-1}, S_{3t-1}$ take values in the set $\{1, 2\}$ (in total, there are eight conditional probability functions, considering all the permutations of $S_{1t-1}, S_{2t-1}, S_{3t-1}$ in the set $\{1, 2\}$). It is interesting to observe the U-shape of these

conditional probability functions. This means that when the three markets were in decline, it is more likely in the next period, which the returns of the SP500 will be in the lowest or highest percentiles, but not in the middle ones (i.e. representing the moderate values of the process).

4. Conclusions

We propose a new method to estimate MMCs of order one or higher. Through a numerical analysis, a Monte Carlo experiment and an empirical application, we have shown that the proposed method is more precise than the MTD model.

Our model can be easily adjusted to model higher-order Markov chain. To illustrate this point, suppose that S_{1t} depends on S_{1t-1}, S_{1t-2} and $S_{2,t-1}$. Then, according to our model, $P_1^{\Phi}(i_0|i_1,\ldots,i_s)$ may be written as

$$\frac{\Phi(\eta_{10}+\eta_{11}P(S_{1t}=i_0|S_{1,t-1}=i_1)+\eta_{12}P(S_{1t}=i_0|S_{1,t-2}=i_2)+\eta_{13}P(S_{1t}=i_0|S_{2,t-1}=i_3))}{\Sigma},$$

where Σ is the normalizing constant (as described before).

The empirical application illustrated the potential use of MMC models. In particular, the results suggest that the model may be able to generate trading rules. This is an issue that may be worth analysing in a future paper. There are several other aspects that can be exploited. In fact, because 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 MMC . For example, using the expression P_j^{Φ} and the estimates $\hat{\eta}_{jk}$, we may compute the conditional mean and volatility over time as follows

$$\hat{\mu}_t = \sum_{k=1}^{10} m_k \times \hat{P}_1^{\Phi}(i\,|S_{1t-1},S_{2t-1},S_{3t-1}),$$

$$\hat{\sigma}_t^2 = \sum_{k=1}^{10} m_k^2 \times \hat{P}_1^{\Phi}(i\,|S_{1t-1},S_{2t-1},S_{3t-1}) - \hat{\mu}_t^2,$$

where m_k is a representative value of the kth class interval $\left[\hat{q}_{(k-1)/100},\hat{q}_{k/100}\right]$ (e.g. the midpoint).

Acknowledgements

This research was supported by the Fundação para a Ciência e a Tecnologia. I am thankful to the referees for their insightful and constructive comments.

References

Adke, S. & Deshmukh. (1988). Limit distributions of a high order Markov chain. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **50**, 105–108.

Berchtold, A. (1996). Modélisation autorégressive des chaî nes de Markov: utilisation d'une Matrice Différente pour Chaque Retard. *Revue de Statistique Appliquée* **44**, 5–25.

Berchtold, A. (1998). *Chaînes de Markov et modèles de transition: applications aux sciences sociales*, HERMES, Paris.

Berchtold, A. (2001). Estimation in the mixture transition distribution model. *J. Time Series Anal.* **22**, 379–397.

Berchtold, A. & Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statist. Sci.* **17**, 328–356.

Chen, D. & Lio, Y. (2009). A novel estimation approach for mixture transition distribution model in high-order Markov chains. *Comm. Statist. Simulation Comput.* **38**, 990–1003.

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

Ching, W., Fung, E. & Ng, M. (2004). Higher-order Markov chain models for categorical data sequences. *Naval Res. Logist.* **51**, 557–574.
Ching, W., Ng, M. & Fung, E. (2008). Higher-order multivariate Markov chains and their applications. *Linear Algebra Appl.* **428**, 492–507.
Le, N., Martin, R. & Raftery, A. (1996). Modelling flat stretches bursts and outliers in time series using mixture transition distribution models. *J. Amer. Statist. Assoc.* **91**, 1504–1515.
Lèbre, S. & Bourguignon, P. (2008). An EM algorithm for estimation in the mixture transition distribution model. *J. Stat. Comput. Simul.* **78**, 713–729.
MacDonald, I. & Zucchini, W. (1997). *Hidden Markov and other models for discrete valued time series*, Chapman & Hall, London.
Madsen, K., Nielsen, H. & Tingleff, O. (2004). *Methods for non-linear least squares problems*, Technical University of Denmark, Lyngby, Denmark.
Mehran, F. (1989a). *Longitudinal analysis of employment and unemployment based on matched rotation samples*, Report, International Labour Office, Bureau of Statistics, Geneva.
Mehran, F. (1989b). Analysis of discrete longitudinal data: infinite-lag Markov models. In *Statistical data analysis and inference* (ed Dodge, Y.), Elsevier Science Publishers, North-Holland, Amster; 533–541.
McQueen, G. & Thorley, S. (1991). Are stock returns predictable? A test using Markov chains. *J. Financ.* **46**, 239–263.
Raftery, A. (1985a). A model for high-order Markov chains. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **47**, 528–539.
Raftery, A. (1985b). A new model for discrete-valued time series autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni* **3-4**, 149–162.
Raftery, A. (1993). Change point and change curve modeling in stochastic processes and spatial statistics. *J. Appl. Statist. Sci.* **1**, 403–423.
Raftery, A. & Banfield, J. (1991). Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 32–43.
Raftery, A. & Tavaré, S. (1994). Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution (MTD) model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **43**, 179–199.
Zhu, D. & Ching, W. (2010). A new estimation method for multivariate Markov chain model with application in demand predictions. In *BIFE'10 Proceedings of the 2010 Third International Conference on Business Intelligence and Financial Engineering* (eds Yu, L., Lai, K. & Wang, S.), IEEE, Computer Society, Hong Kong.

Received January 2013, in final form November 2013

João Nicolau, ISEG, Universidade de Lisboa, Rua do Quelhas 6, 1200-781 Lisboa, Portugal.
E-mail: nicolau@iseg.utl.pt

Supporting information

Additional information for this article is available online including the data and the routines in GAUSS to estimate the models (also available at site: http://pascal.iseg.utl.pt/~nicolau/myHP/codes.rar).