Universidade de Lisboa

Instituto Superior de Economia e Gestão

Mestrado em Econometria Aplicada e Previsão

# Multivariate Markov Chains - Estimation, Inference and Forecast. A New Approach: What If We Use Them As Stochastic Covariates?

Bruno Miguel Pinto Damásio

Setembro de 2013

Orientação: João Nicolau

***Multivariate Markov Chains - Estimation, Inference and* Forecast*. A New Approach: What if We Use Them as Stochastic Covariates?***

Bruno Damásio

*M.Sc.:* Applied Econometrics and Forecasting

*Supervisor:* João Nicolau

**Abstract**

This dissertation proposes a new concept: the usage of Multivariate Markov Chains (MMC) as covariates. Our innovative approach is based on the observation that we can treat possible categorical regressors as a MMC in order to improve the forecast error of a certain dependent variable, provided it is caused, in the Granger sense, by the MMC. We conduct a Monte Carlo simulation study to assess the performance of our model and we archive excellent results in terms of forecast. An empirical illustration, that widely supports the results obtained in the Monte Carlo study, is also provided. Furthermore, the results of our empirical illustration suggest that the sovereign bond markets in peripherical European countries, namely Portugal, are inefficient. The conclusions drawn include implications for policy. We also discuss the ideas behind several methods to estimate MMC, tackling issues with regard to the statistical inference topic. We provide a general framework to allow us to obtain the MMC h-step-ahead forecast closed formulas.

**Keywords:** Markov chains as covariates, Multivariate Markov chains, High order Markov chains, Mixture transition distribution.

**JEL codes:** C18, C51, C53, C58, G17.

***Multivariate Markov Chains - Estimation, Inference and Forecast. A New Approach: What if We Use Them as Stochastic Covariates?***

Bruno Damásio

*Mestrado:* Econometria Aplicada e Previsão

*Orientação:* João Nicolau

**Resumo**

Esta dissertação propõe um novo conceito: a utilização de Cadeias de Markov Multivariadas enquanto regressores. A nossa abordagem inovadora baseia-se na observação de que é possível fazer uso de CMM enquanto variáveis explicativas com o intuito de se reduzirem os erros de previsão de uma determinada variável dependente, desde que essa variável dependente seja causada, *a la* Granger, pela CMM. Com o objectivo de perceber a performance do nosso modelo em termos de previsão operacionalizamos um estudo de simulação de Monte Carlo no qual obtemos excelentes resultados. Também recorremos a uma ilustração empírica que sustenta fortemente os resultados obtidos no estudo de simulação de Monte Carlo. Para além disso, os resultados da ilustração empírica apontam para a circunstância de que os mercados das obrigações das dívidas soberanas dos países da periferia europeia, nomeadamente Portugal, são ineficientes. Podem retirar-se das conclusões obtidas algumas implicações em termos de orientação de política económica. Discutimos ainda algumas ideias subjacentes às diversas metodologias de estimação de CMM, sublinhando as questões relativas ao tópico da inferência estatística. Providenciamos uma utensilagem teórica do seio da qual se obtêm as expressões da previsão a h-passos com CMM.

**Palavras-chave:** Cadeias de Markov enquanto regressores, Cadeias de Markov multivariadas, Cadeia de Markov de ordem superior, Distribuição de mistura de transições.

**Códigos JEL:** C18, C51, C53, C58, G17.

## Acknowledgments

The guidance of Professor João Nicolau throughout this work was exceptional. I would like to thank him for having been a super supervisor.

# Contents

## List of Figures

# List of Tables

# 1 Introduction

This essay adopts a new approach to the treatment of Multivariate Markov Chains (MMC) as stochastic categorical covariates. It is a relevant innovation since the usage of MMC as regressors is a completely new concept for reasons that have never been raised in the published literature. Our research question is as follows: is it empirically feasible to use MMC as regressors? Will it improve the forecast error of a certain dependent variable?

Raftery (1985) has proposed a method to represent and to estimate high-order dependencies among categorical data sequences: the mixture transitions distribution model (MTD). Ching (2002), has used Raftery's MTD model to estimate dependencies among an interrelated multivariate categorical stochastic process - a MMC. Until then, the estimation of MMC had proven a problematic task.

The next section introduces some preliminary concepts that are the basic toolkit to understand the fundamentals of the MMC model. Section 3 provides a brief literature review of MMC models, stressing the latest breakthroughs in MMC estimation and the basic problems that ensue from such estimations. Section 4 firstly covers the theoretical framework of estimation and statistical inference of the method chosen - the MTD Probit, and secondly discusses the issue of forecasting. In this section we also present the framework and the assumptions under which we can obtain the closed formulas of MMC h-step-ahead forecast. Section 5 presents our innovative concept: the use of MMC as covariates. In fact, we hypothesize that if a MMC plays the role of regressors then we might improve the forecast error of a certain dependent variable, given that it is, in the Granger sense, caused by the MMC. These improvements will be evaluated, first of all, through a Monte Carlo simulation problem. Afterwards, we provide an economic illustration with respect to the relationship between some southern European sovereign credit ratings and Portuguese sovereign bond yields. Section 6 discusses some possible extensions of this essay. Finally, Section 7 elaborates on the summary of the main results and contains some concluding remarks.

## 2 Some Preliminary Concepts

In the early twentieth century Andrey Markov proposed a probabilistic model to typify a pattern of dependencies in a stochastic process. More precisely, a model able to capture intra-probability transitions, with respect to past events, within categorical data sequences. Mathematically, for the discrete time case, one may conclude the Markovian property states that given the present, the future is independent of the past, as follows:

$$P(X_t = a | \mathcal{F}_{t-1}) = P(X_t = a | X_{t-1}) \tag{2.1}$$

where $\mathcal{F}_{t-1}$ is the $\sigma - algebra$ generated by the available information until $t-1$. When the events (past, present and future) represent a specific state, we get a Markov chain (MC) - a Markov process defined into a countable (finite or infinite) state space set $E = \{1, \cdots, m\}$ or $E = \{1, 2, \cdots\}$. Given a $m \times m$ matrix - the one step probability transitions matrix (PTM) and the initial conditions, we can fully characterize a MC. Each PTM row represents a probability function - adding up to one and are non-negative. Below we illustrate a PTM.

$$\begin{bmatrix} P(X_t = 1 | X_{t-1} = 1) & \cdots & P(X_t = m | X_{t-1} = 1) \\ \vdots & \ddots & \vdots \\ P(X_t = 1 | X_{t-1} = m) & \cdots & P(X_t = m | X_{t-1} = m) \end{bmatrix} \tag{2.2}$$

In some applications, in order to facilitate the implementation of some theorems and results it might be important to think in terms of long-term probability events. Formally, it might be interesting to evaluate:

$$\lim_{h \to \infty} P(X_{t+h} = a | \mathcal{F}_{t-1}) \tag{2.3}$$

The following proposition guarantees the existence of the previous limit.

**Proposition 2.1.** *Ergodicity*

*A MC is said to be ergodic if it is positive recurrent and aperiodic. Under those circumstances the row vector of stationary probabilities* $\boldsymbol{\pi} \equiv [\pi_1, \cdots, \pi_m]$ *exists and satisfies the following equation:*

$$\boldsymbol{\pi} P = \boldsymbol{\pi}, \ with \ \sum_{i=1}^{m} \pi_i = 1 \ and \ \pi_i \geq 0 \tag{2.4}$$

*where* $P$ *is the PTM associated with the MC.*

$$\pi_i = \lim_{h \to \infty} P(X_{t+h} = i | X_{t-1}) \tag{2.5}$$

Proposition (2.1) states that a sufficient condition for the existence, and for the uniqueness, of the MC stationary representation is, on the one hand, that each state communicates with each other, which implies that there are not absorbent states (if the process yesterday was in state $i$ then it will return to $i$, in a finite time-horizon, with probability one) and, on the other hand, the fact that no-one is preventing that the state $i$ is revisited in two consecutive moments. In practical terms, we may state that a MC is ergodic if and only if it is possible to go to every state from every state, of course, not necessarily in just one step. We will address this issue later in the context of Multivariate Markov Chains (MMC). The question is: what is a MMC?

Suppose, for now, that we have $s > 1$ categorical time series (categories) interrelated. When the state of the future events of a category depends not only on its previous state (inter-transition) but also on another series' previous states (intra-transitions) we get a MMC. MMC plays an important role and is a valuable toolkit for working on various topics in several science subjects, such as credit and financial data modeling, economics, biology, history, meteorology, chemistry, sociology, music and linguistics, among many other topics (Berchtold and Raftetry, 2002).

Introducing some notation and some concepts, formally, we assume that we observe a realization of a multivariate discrete stochastic Markov process $\{(S_{1t}, \cdots, S_{st})\}$ for $T$ observations $(t = 1, \cdots, T)$ where each $S_{jt}$ can take values in the countable set $E = \{1, \cdots, m\}$ and $j = 1, \cdots, s$. Without any loss of generality, we also assume that we have a first order MMC, in the sense that

$$P\left(S_{jt} = k \middle| \mathcal{F}_{t-1}\right) = P\left(S_{jt} = k \middle| S_{1t-1} = i_1, \cdots, S_{st-1} = i_s\right) \tag{2.6}$$

Therefore, one assumes that in order to explain and forecast $S_{jt+1}$ the past values of the process, $S_{jt-i,}$ $i > 0$, are needless since we require only its present values. The assumption that arises from equation (2.6) is not a constraint, not even a mild restriction. Indeed, as we will see in the next sections:

1. High order MC (HOMC) can be viewed as a particular case of MMC.
2. We can estimate them even as one can estimate high-order MMC (HMMC).
3. We can use those estimates to forecast within the MMC and we can also use them to help forecast a certain dependent variable.

# 3 Review of Multivariate Markov Chains Models

One might say that the HOMC was the genesis for the MMC, as we shall see in further detail. Accordingly, Raftery (1985) introduced the mixture transition distribution model (MTD) as an appropriate model to represent high-order dependencies within a data sequence. A MMC, or, roughly, a HOMC, involves $m^s$ states (where $m$ represents the number of states and $s$ denotes the number of categorical series). Therefore, this represents the main problem regarding the conventional MMC, i.e. of facing a MMC as an usual MC model. This might be a problematic issue for several reasons: 1) the number of parameters is huge - $m^{s+1}$, which can render the estimation a daunting task 2) the size of the transitions matrix is also very large 3) it is a very hard task, even using computational brute force, to find the stationary probability vector, 4) the parameters cannot be efficiently estimated (as the standard errors present an explosive behavior), 5) in finite samples, the parameters may not even be identified. Table (3.1) displays the number of parameters of the usual MMC model as a function of the number of states and of the number of categorical series, respectively, $m$ and $s$. An application involving 5 categorical series with a space state of 10 elements consists of more than 48 million parameters. If we add one more categorical series then the overall number of parameters of the model rises to 362 million parameters.

Due to this obstacle, Raftery (1985) argued that MTD was more suitable than some competing high-order MC models at that time, such as Jacobs and Lewis (1978), Pegram (1980) and Logan (1981), both in terms of adjustment criteria, like AIC, and in terms of parsimony (since it involves less unknown parameters). The author illustrated the method through three MMC empirical applications. We can represent the MTD model as follows:

$$P\left(X_t = i_o \middle| X_{t-1} = i_1, \cdots, X_{t-l} = i_l\right) = \sum_{g=1}^{l} \lambda_g P\left(X_t = i_o \middle| X_{t-g} = i_g\right), \, t = 1, ..., T; \, g = 1, ..., l. \tag{3.1}$$

To ensure that the quantities

$$P\left(X_t = i_o \middle| X_{t-1} = i_1, \cdots, X_{t-l} = i_l\right) \tag{3.2}$$

**Table 3.1:** Usual MMC model: Number of Parameters

| $s$ | $m$ | $n$ |
|---|---|---|
| 4 | 5 | 4.096 |
| 5 | 6 | 78.125 |
| 5 | 10 | 48.828.125 |
| 6 | 10 | 362.797.056 |

are probabilities, i.e. that they are non-negative and less than or equal to 1, one may assume that expression (3.2) is a linear convex combination of the components

$$P\left(X_t = i_o \middle| X_{t-g} = i_g\right) \tag{3.3}$$

by imposing the following restrictions:

$$\sum_{g=1}^{l} \lambda_g = 1 \tag{3.4}$$

$$\lambda_g \geq 0 \tag{3.5}$$

As we will see later, restriction (3.5) has a practical implication: it assumes that the categorical series are positively correlated. Nevertheless, it is a sufficient but not necessary condition to ensure that the probability terms (3.2) are non-negative and less than one. In fact, a less restrictive condition, together with restriction (3.4) is that

$$0 \leq \sum_{g=1}^{l} \lambda_g P\left(X_t = i_o \middle| X_{t-g} = i_g\right) \leq 1 \tag{3.6}$$

The benefits of assuming the positivity condition (condition 3.5) is that the estimation becomes simpler and the $\lambda_g$ parameters can be seen as probabilities (Raftery and Tavaré 1994). We might state that there are two special cases of the MTD model: with and without assuming the positivity condition. We shall discuss this further later.

The MTD model has been used in several applications in manifold scientific fields. Berchtold and Raftery (2002) presents a comprehensive survey on MTD. It reviews the MTD model and analyzes some major development issues from 1985 to 2002, such as MTD parameters estimation, presenting many MTD applications and generalizations, and it illustrates some other possible ways to estimate high-order MC, addressing some inference issues regarding the MTD model.

As for recent years, since 2002, no systematic surveys have focused on the problem. We can distinguish two main approaches: the Raftery HOMC followers and the Ching (2002) MMC followers[1]. As Ching is, himself, a Raftery follower, one might say that Raftery is the father of both theoretical frameworks: HOMC and MMC concepts. Let us now focus on a few points regarding the subject.

With regard to Raftery's MTD followers, without going into many technicalities, we highlight Berchtold (2001) who proposes a new iterative algorithm for MTD estimation, concluding that this method performs at least as well as the competing methods. Lèbre and Bourguignon (2008) propose an Expectation-Maximization algorithm, which is easier to use than that of Berchtold. Lastly, Chen and Lio (2009) propose a novel

[1]Another class of HOMC models is the Polytomus (logistic) regression models, where algebraically we have $\ln \frac{P(X_t=1|X_{t-1},\cdots,X_{t-l})}{P(X_t=0|X_{t-1},\cdots,X_{t-l})} = \beta_o + \sum \beta_l X_{t-l} + u_t$. See Rajarshin (2013), Kvam and Sokol (2006), Wasserman and Pattinson (1996) and Azzalini (1993). We will not address this issue here.

approach of MLE, converting the nonlinear embedded constraints into box constraints. With respect to Raftery's MTD generalizations, Berchtold and Raftery (2002) discuss some relevant extensions to the MTD model. The first one is the Multimatrix MTD, Berchtold (1995, 1996, 1998). The original MTD uses the same TPM to model the dependencies between present and each lag term. Here is proposed to relax this assumption by using a different $m \times m$ transition matrix for each lag.

$$P\left(X_t = i_o \middle| X_{t-1} = i_1, \cdots, X_{t-l} = i_l\right) = \sum_{g=1}^{l} \lambda_g P\left(X_t = i_o \middle| X_{t-g} = i_g\right)^{(g)} \tag{3.7}$$

Another possible generalization is the Infinite-Lag MTD model that assumes an infinite lag order - $l = \infty$ as in Mehran (1998), Le, Martin and Raftery (1996). The third generalization allows the modeling of data sets with missing data: The missing data MTD model, for instance, assumes the sequence:

$$\{X_1, X_2, \cdots, X_{t-k-1}, ?, X_{t-k+1}, \cdots, X_t\} \tag{3.8}$$

where the $k - th$ entry - $X_k$ is unknown. Finally, another relevant generalization is the MTD with General State Spaces: which allows to model more general processes with an arbitrary space state as in Martin and Raftery (1987), Adke and Deshmukh (1988), Wong and Li (2000). One assumes constraints (3.5) and (3.4) plus the following parametric specification:

$$F\left(X_t \middle| \mathcal{F}_{t-1}\right) = \sum_{g=1}^{l} \lambda_g G_g\left(X_t \middle| X_{t-g}\right) \tag{3.9}$$

- $F\left(X_t \middle| \mathcal{F}_{t-1}\right)$ is the conditional distribution of $X_t$
- $G_g\left(X_t \middle| X_{t-g}\right)$ is an arbitrary cumulative distribution function (cdf)

Once set $G_g\left(X_t \middle| X_{t-g}\right)$ as the standard Normal distribution, that is

$$G_g\left(X_t \middle| X_{t-g}\right) = \Phi\left(\frac{X_t - \phi_g X_{t-g}}{\sigma_g}\right) \tag{3.10}$$

we obtain the Gaussian MTD model (GMTD)[2], as presented in Le, Martin and Raftery (1996). For more details and for more generalizations see Berchtold and Raftery (2002).

As stated, one can see a HOMC as a MMC. This assertion is due to the work of Ching et al (2002) who, using the MTD model, conceptualized the HOMC model as a particular case of the MMC, therefore generalizing the concept of the MTD model. In fact, until 2002 there are few studies tackling the MMC estimation issue[3].

[2] The expression (3.10) may be seen as a particular case of a regime-switching model with independent states.

[3] See, for instance, Gottschau (1992). The main problem of these studies is that they become unfeasible when we have a large number of states and/or categorical series.

Unlike the univariate methods (even high-order methods) which only enable the capturing of intra-probability transitions (within sequences), the greatest merit of the MMC model is that it allows intra and inter-probability transitions within and between categorical data sequences to be captured.

The method considered by Ching is, in fact, a pseudo-generalization of Raftery's MTD, since the innovation is just conceptual (HOMC as a MMC): the model is the same, a MTD with the positivity assumption (assumes 3.5). Ching applies the positivity version of MTD but, unlike Raftery, to the MMC case. One can say that Ching's work was of great importance for two reasons: 1) it led high-order univariate MC to be viewed and conceptualized as a MMC for the first time, and 2) in the last 10 years the majority of the published articles on MMC follow Ching's concept. To better understand it, let us consider the categorical data sequence $\{(S_{1t,\cdots,}S_{st})\}$, defined in previous section, having $m$ states. We rewrite the process $\{S_{jt}\}$, using the $m-row$ standard basis vectors, as the state vector sequences:

$$\left\{\boldsymbol{x}_t^{(1)},\cdots,\boldsymbol{x}_t^{(s)}\right\} \tag{3.11}$$

Where,

$$\boldsymbol{x}_t^{(j)} = \begin{cases} \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}' & \text{if} \quad S_{jt}=1 \\ \begin{bmatrix} 0 & 0 & \cdots & \underbrace{1}_{k-th\ entry} & \cdots & 0 \end{bmatrix}' & \text{if} \quad S_{jt}=k \\ \begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}' & \text{if} \quad S_{jt}=m \end{cases} \tag{3.12}$$

$$\boldsymbol{x}_{t+1}^{(j)} \approx \sum_{k=1}^{s} \lambda_{jk} P^{(jk)} \boldsymbol{x}_t^{(k)},\ \text{for } j=1,\cdots,s \tag{3.13}$$

The $m \times m$ matrices $P^{(jk)}$ have as a generic $vu$ element the scalar:

$$P_{uv}^{(jk)} \equiv P\left(S_{jt}=u \middle| S_{kt-1}=v\right) \tag{3.14}$$

These elements may be estimated through the maximum likelihood method:

$$\hat{P}\left(S_{jt}=u \middle| S_{kt-1}=v\right) = \frac{n_{vu}}{\sum_{u=1}^{m} n_{vu}} \tag{3.15}$$

where $n_{vu}$ represents the number of transitions to $S_{jt}=u$ from $S_{kt-1}=v$. Then, we have, likewise the more restrictive version of Raftery's MTD model - with the positivity assumption, a linear convex combination between the different components, and we also have

$$0 \leq \lambda_{jk} \leq 1 \text{ with } 1 \leq j,k \leq s \text{ and} \sum_{k=1}^{s} \lambda_{jk} = 1 \tag{3.16}$$

Writing the model in matrix form, and assuming the equality in equation (3.13) we have:

$$\underline{\boldsymbol{x}}_{t+1} \equiv \begin{bmatrix} \boldsymbol{x}_{t+1}^{(1)} \\ \vdots \\ \boldsymbol{x}_{t+1}^{(s)} \end{bmatrix} = \begin{bmatrix} \lambda_{11}P^{(11)} & \cdots & \lambda_{1s}P^{(1s)} \\ \vdots & \ddots & \vdots \\ \lambda_{s1}P^{(s1)} & \cdots & \lambda_{ss}P^{(ss)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t}^{(1)} \\ \vdots \\ \boldsymbol{x}_{t}^{(s)} \end{bmatrix} \equiv Q\underline{\boldsymbol{x}}_t \tag{3.17}$$

Where $Q$ is an $ms \times ms$ block matrix ($s \times s$ blocks of $m \times m$ matrices) and $\boldsymbol{x}_t$ is a stacked $ms$ column vector ($s$ vectors, each one with $m$ rows) . Expression (3.13) has a practical implication: we can see that the state probability distribution of the $j-th$ sequence depends on $\sum_{k=1}^{s} \lambda_{jk} P^{(jk)} \boldsymbol{x}_t^{(k)}$ and, since we have (3.16), this corresponds to the weighted average of the terms $P^{(jk)}\boldsymbol{x}_t^{(k)}$. Consequently, to obtain the quantity $\boldsymbol{x}_{t+1}^{(j)}$ we just need to estimate the matrices $P^{(jk)}$ and the quantities $\lambda_{jk}$.

Regarding the latter, the $\lambda_{jk}$ coefficients, the estimation method proposed by Ching involves the following optimization problem:

$$\begin{cases} & min_{\lambda} max_i \left| \left[ \sum_{k=1}^{m} \lambda_{jk} \hat{P}^{(jk)} \hat{\mathbf{x}}^{(k)} - \hat{\mathbf{x}}^{(j)} \right] \right| \\ s.t. & \sum_{k=1}^{s} \lambda_{jk} = 1 \ and \ \lambda_{jk} \geq 0 \end{cases} \tag{3.18}$$

As we will show next, 1) the matrices $P^{(jk)}$, or, should we say, their consistent estimator $\hat{P}^{(jk)}$, are of the utmost importance because, as far as we know, all the estimation methods share them, even though they differ with regard to the estimation of the remaining parameters and 2) the method considered by Ching is highly inefficient.

Despite the obvious similarities between the functional forms of Ching's and Raftery's models, it is important to emphasize the completely different ways proposed to estimate the unknown parameters, since Raftery employs the maximum likelihood method (MLE) to estimate them. Another difference concerns the number of unknown parameters. This happens because Ching's model has more equations than Raftery's model by reason of it is completely nonsense[4] to evaluate, for instance, lead probabilities like:

$$P\left(S_{jt-1} = i_o \middle| S_{kt} = i_1\right) \tag{3.19}$$

However, this is not a relevant issue because the estimation may be carried out equation by equation, and, within equations the two models share the same number of parameters.

By writing Ching's MMC model using Raftery's HOMC model notation, we obtain our MTD model for MMC:

[4]It may make sense in reversibility time series fields, but that falls outside the scope of this paper.

$$P\left(S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s)^{MTD}=$$
$$\lambda_{j1}P\left(S_{jt}=i_o\right|S_{1t-1}=i_1)+\cdots+\lambda_{js}P\left(S_{jt}=i_o\right|S_{st-1}=i_s)=$$
$$\sum_{k=1}^{s}\lambda_{jk}P\left(S_{jt}=i_o\right|S_{kt-1}=i_k) \quad (3.20)$$

Consequently, a HOMC is a MMC in the sense that we can specify each categorical series as follows:

$$\begin{aligned} S_{1t} &= X_t \\ S_{2t} &= X_{t-1} \\ &\vdots \\ S_{st} &= X_{t-s+1} \end{aligned} \quad (3.21)$$

From the beginning of the $21^{st}$ century, in particular since 2002 onwards, a lot of scientific articles on the subject have been published. Although there is a lot of research on the MMC theme, there is not much disparity in the models used in the published papers, since most of the papers employ either the model considered by Ching et al (2002) or a consequent generalization. For instance, Fung et al (2002) employs it in the construction of a wind turbine in a certain wind farm by analyzing the wind speed form several potential locations, Oz and Erpolat (2011) applies Ching's original model to fluctuations in the euro and dollar exchange rates against the Turkish lira and Liu (2010) analyzes and predicts price and sales volume of a certain product. On the other hand, we can speculate that over the last few years the published studies have undergone major improvements in terms of parsimony by reducing the number of unknown parameters of the model, originally $s^2\left(m^2+1\right)$. In fact, Kijima et al (2002) proposed a parsimonious MMC model to simulate correlated credit risks and Siu et al (2005), on the same issue, proposed a less parsimonious model but one that was easier to manipulate than that of Kijima et al (2002), with $s^2m^2$ parameters. Zhang et al (2006) develops a simplified model, albeit tending towards the model of Ching et al (2002) where the number of parameters is reduced to $s\left(m^2+1\right)$. It proposes a simple assumption: $P^{(jk)}=I_m$ when $j\neq k$. The $Q$ matrix, defined in (3.17) becomes

$$Q=\begin{bmatrix} \lambda_{11}P^{(11)} & \lambda_{12}I_m & \cdots & \lambda_{1s}I_m \\ \lambda_{21}I_m & \lambda_{22}P^{(22)} & & \lambda_{2s}I_m \\ \vdots & & \ddots & \vdots \\ \lambda_{s1}I_m & \lambda_{s2}I_m & \cdots & \lambda_{ss}P^{(ss)} \end{bmatrix} \quad (3.22)$$

This simplification has a practical implication: one the one hand, if $S_{kt-1} = u$ then $S_{jt} = v$, for $k \neq j$ and for $u \neq v$ with probability zero, on the other hand $S_{kt-1} = u$ then $S_{jt} = v$, for $k \neq j$ and for $u = v$ with probability one. We have, for $k \neq j$ :

$$P\left(S_{jt} = v \middle| S_{kt-1} = u\right) = \begin{cases} 0 & \text{if } u \neq v \\ 1 & \text{if } u = v \end{cases} \tag{3.23}$$

Ching et al (2007a) considered the assumption on the $Q$ block matrix proposed in Zhang et al (2006) to forecast sales demand data sequences, and proposed a simplified model to overcome the two main problems of dealing with very short data length structures. Namely, very large forecast errors on the transition probability matrices and unreached steady-states. While Zhang et al (2008) uses the simplification considered in Zhang et al (2006) to approximate a Probabilistic Boolean Network in order to control genetic regulatory networks.

Regarding more efficiency improvements, Zhu and Ching (2010) propose a new method of estimation based on the minimization of the unconditional forecast error mean, involving some nonlinear programming problems. Nicolau (2012) translates the original problem of Zhu and Ching (2010), which involves inequality restrictions, into an unrestricted non-linear least squares estimation. In fact, Nicolau, as well as Zhu and Ching, noticed that the method proposed by Ching et al (2002) is not optimal, in the sense that it is based on the unconditional mean and, as is it known, the best predictor (in mean squared error) is the conditional mean.

Lastly, another important topic is related to the fact that, in order to ensure that we have probabilities, Ching and his followers assume a convex combination between the terms and impose the restrictions (3.16). All these studies say usual MMC models share a common denominator: they assume a positive correlation between the different data sequences due to the restrictions (3.16). This assertion implies that if, at the moment $t$, one of the sequences, for instance, $S_{jt}$ has an increase in its state probability, then it can only have an increasing, and never a decreasing, impact in the state probability of $S_{kt+1}$ for $k \neq j$.

Consequently, it can easily be shown that if we have a negative correlation between series, for instance, $Corr(S_{j.}, S_{k.}) < 0$, the quantities $\lambda_{jk}$ are forced to be zero. Furthermore, as is well known, correlation and causality are very different concepts - a correlation relationship between two random variables, $A$ and $B$, does not necessarily imply that one causes another. For instance, given a third variable, $C$, that, by assumption, causes both $A$ and $B$ the conditional correlation, on $C$, between $A$ and $B$ can be zero despite the fact that the marginal correlation is positive. So, even if we have a positive correlation between variables, say $S_{1t}$ and $S_{2t-1}$, controlling for a third common effect between two random variables, $S_{3t-1}$, we can have negative conditional correlations, in other words, given $S_{3t-1}$, $S_{1t}$ and $S_{2t-1}$ may be negatively correlated sequences. This is another feature that the standard positivity MMC models (common MMC models) cannot capture: positive correlations but negative conditional correlations, or, perhaps we should say, negative causality relationships.

Another problem shared by common MMC models (MMC models with positivity assumption - *a la* Ching) is that, as they are grounded on the said convex combination, another restriction is imposed:

$$\begin{aligned} &Min\left\{P\left(S_{jt}=i_o\right|S_{kt-1}=i_k)\right\} \leq \\ &\qquad P\left(S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s) \\ &\qquad\qquad \leq Max\left\{P\left(S_{jt}=i_o\right|S_{kt-1}=i_k)\right\}\ for\, k=1,\cdots,s \end{aligned} \tag{3.24}$$

The term $P\left(S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s)$ is bounded between the minimum and the maximum of $P\left(S_{jt}=i_o\right|S_{kt-1}=i_k)$.

Since the usual MMC models are based on the assumption (3.16), an obvious solution to relax the previous assumptions is not to assume the constraints (3.16), without assuming anything else. The problem is that the results produced by the model are no longer probabilities. Several solutions able to deal with the aforementioned problems have been provided. Raftery and Tavaré (1994) developed a strategy, dropping the positivity condition, by imposing a new set of restrictions

$$Tq_i^- + (1-T)\,q_i^+ \geq 0 \tag{3.25}$$

where

$$T = \sum_{\substack{g=1 \\ \lambda_g \geq 0}}^{m} \lambda_g \tag{3.26}$$

$$q_i^- = min_{1\geq j\geq m} q_{ij} \tag{3.27}$$

$$q_i^+ = max_{1\geq j\geq m} q_{ij} \tag{3.28}$$

$$q_{ij} = P\left(S_{kt}=j\right|S_{lt-1}=i) \tag{3.29}$$

Raftery and Tavaré (1994) have shown that the restriction (3.25) is equivalent to the expression (3.6).

Ching et al (2007b) inspired by Raftery and Tavaré (1994) and on the Zhang et al (2006) $Q$ matrix simplification, propose the following idea to handle with negative correlations between the state vector $\boldsymbol{z}_{t+1}$ and $\boldsymbol{x}_t$:

$$\underline{\boldsymbol{z}}_{t+1} = \frac{1}{m-1}\left(\underline{\mathbf{1}}_{ms} - \underline{\boldsymbol{x}}_t\right) \tag{3.30}$$

where $\mathbf{1}_{ms}$ is $ms$ stacked vector of ones. Then, they split the $Q$ matrix into the sum of two other matrices, where one represents the positive correlations and another the negative correlations, as follows:

$$\begin{bmatrix} \boldsymbol{x}_{t+1}^{(1)} \\ \vdots \\ \boldsymbol{x}_{t+1}^{(s)} \end{bmatrix} = \Lambda^{+} \begin{bmatrix} \boldsymbol{x}_{t}^{(1)} \\ \vdots \\ \boldsymbol{x}_{t}^{(s)} \end{bmatrix} + \frac{1}{m-1}\Lambda^{-} \begin{bmatrix} \mathbf{1}_m - \boldsymbol{x}_{n}^{(1)} \\ \vdots \\ \mathbf{1}_m - \boldsymbol{x}_{n}^{(s)} \end{bmatrix} \tag{3.31}$$

notice that

$$\Lambda^{+} \equiv \begin{bmatrix} \lambda_{11}P^{(11)} & \lambda_{12}I_m & \cdots & \lambda_{1s}I_m \\ \lambda_{21}I_m & \lambda_{22}P^{(22)} & & \lambda_{2s}I_m \\ \vdots & & \ddots & \vdots \\ \lambda_{s1}I_m & \lambda_{s2}I_m & \cdots & \lambda_{ss}P^{(ss)} \end{bmatrix} \tag{3.32}$$

and

$$\Lambda^{+} \equiv \begin{bmatrix} \lambda_{1-1}P^{(11)} & \lambda_{1-2}I_m & \cdots & \lambda_{1-s}I_m \\ \lambda_{2-1}I_m & \lambda_{2-2}P^{(22)} & & \lambda_{2-s}I_m \\ \vdots & & \ddots & \vdots \\ \lambda_{s-1}I_m & \lambda_{s-2}I_m & \cdots & \lambda_{s-s}P^{(ss)} \end{bmatrix} \tag{3.33}$$

They have applied the model above on a sales demand forecast application. This model, moreover, was successfully applied in Ching et al (2009) through two credit risk applications and Wang and Huang (2013) tested the convergence conditions of the model through numerical experiments. However, the restrictions (3.16) were maintained in all those models.

Nicolau (2013) has proposed a completely different way to deal with the problems without assuming (3.16), consequently, without the restriction (3.24) and also without splitting the $Q$ matrix. The method estimates the unknown parameters through the MLE so, like the concrete dichotomy between Nicolau (2012) and Ching (2002) at least asymptotically it is a better method, mainly in terms of efficiency, than Ching et al's (2009) since it is based on the marginal mean. This is the main caveat of Ching's work (and of its followers): it is highly inefficient. Moreover, Nicolau (2013) has made a solid bridge between Raftery's and Ching's work. Indeed, Nicolau proposes a generalization of Raftery's MTD on the MMC model, like Ching, but estimating it like Raftery: using the MLE method. As is well known, under some regularity conditions, the MLE is asymptotically the best estimator with regard to efficiency. We will elaborate on the problem in the next section.

In a word: Ching (2002) was the first person to use the MTD model applied to the MMC case. Until then, all MTD models had been conducted only on HOMC case. While the original MTD model is estimated through the MLE method, Ching (and his followers) estimated the unknown parameters using an inefficient method. It is clear that the best way to estimate MMC is through the MLE. Nicolau (2013) proposed a generalization of the MTD model to estimate the parameters, which is even better than the original MTD model.

# 4 Multivariate Markov Chains: Theoretical Framework

## 4.1 Multivariate Markov Chains: Estimation and Inference

Here we present the model proposed in Nicolau (2013), discussing some issues concerning the estimation and the inference of MMC.

The prime idea is to model the quantity $P\left(\left.S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s\right)$ as follows:

$$P\left(\left.S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s\right)^{\Phi}\equiv \frac{G\left[\eta_{j0}+\eta_{j1}P\left(\left.S_{jt}=i_o\right|S_{1t-1}=i_1\right)+\cdots+\eta_{js}P\left(\left.S_{jt}=i_o\right|S_{st-1}=i_s\right)\right]}{\sum_{k=1}^{m}G\left[\eta_{j0}+\eta_{j1}P\left(\left.S_{jt}=k\right|S_{1t-1}=i_1\right)+\cdots+\eta_{js}P\left(\left.S_{jt}=k\right|S_{st-1}=i_s\right)\right]} \quad (4.1)$$

Where $G(\cdot)$ may be an arbitrary cumulative probability distribution function (cdf). Without any loss of generality we can use the cdf of the standard normal distribution - $\Phi(\cdot)$, as in Nicolau (2013). The model is called a MTD-Probit due to the fact that it being founded on a cdf and the argument of $G(\cdot)$ is a linear combination of

$$P\left(\left.S_{jt}=i_0\right|S_{kt-1}=i_k\right),\ k=1,\cdots,s \quad (4.2)$$

as in a MTD model.

Despite the obvious advantages of MTD-Probit against traditional MMC models: 1) it is much more efficient, 2) it can capture marginal and conditional negative correlations (the constraints on the $\lambda_{kj}$ are needless) and 3) the restrictions (3.24) are also useless, since the MTD-Probit model results

$$P\left(\left.S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s\right)^{\Phi} \quad (4.3)$$

are bounded between 0 and 1. They are greater than 0 due the numerator of (4.1) and are smaller than 1 due the numerator, and compared to the MTD model: 1) the absence of constraints makes it easier to carry out the standard numerical optimization routines and the MTD-Probit model can generate a larger range of patterns than the MTD model (Nicolau 2013) and 2) the model includes a constant term which may substantially improve the fit of the model. The following theorem, however, is less obvious.

**Theorem 4.1.** MTD-Probit *model probabilities shall be no worse than MTD model probabilities.*

*Let, respectively, the quantities*

$P_j(i_o|\, i_1, \cdots, i_s)$, $P_j(i_o|\, i_1, \cdots, i_s)^{\Phi}$ *and* $P_j(i_o|\, i_1, \cdots, i_s)^{MTD}$ *be the true unknown probability, the MTD-Probit model probability and the MTD model probability of the generic event* $S_{jt} = i_o, S_{1t-1} = i_1, \cdots, S_{st-1} = i_s$.

*Suppose that the rows of transition probability matrices between* $S_{jt}$ *and* $S_{lt}$ *are all dissimilar, so* $S_{jt}$ *and* $S_{lt}$ *are correlated sequences,* $j, l = 1, \cdots, s$. *Therefore, one always has:*

$$min_{\eta_{ji}} \sum_{i_1 \cdots i_s i_o} \left| P_j(i_o|\, i_1, \cdots, i_s) - P_j(i_o|\, i_1, \cdots, i_s)^{\Phi} \right|^2 \leq$$
$$min_{\sum \lambda_{ij}=1} \sum_{i_1 \cdots i_s i_o} \left| P_j(i_o|\, i_1, \cdots, i_s) - P_j(i_o|\, i_1, \cdots, i_s)^{MTD} \right|^2 \quad (4.4)$$

*Proof.* See Nicolau (2013). □

A transition probability matrix is said to be positively regular if and only if any power of the TPM only has positive elements. Adke and Desmukh (1988) concluded that if the TPM is positively regular, the HOMC model admits a unique stationary distribution, hence the HOMC is ergodic. The previous fact can be extended to the MMC model case.

**Theorem 4.2.** *Ergodicity of a MMC*

*If all matrices* $P^{(jk)}$ *defined in (3.14) are positively regular, then the associated MMC is ergodic*

*Proof.* See Adke and Desmukh (1988). □

This outcome allows us to establish the MLE properties for the MTD-Probit model. It can be easily shown that our MTD-Probit MMC model satisfies the Cramer regularity conditions. Moreover, assuming a compact parameter set (i.e. that any plausible value for is $\boldsymbol{\eta}_j$ finite) and the remaining conditions of propositions 7.1 and 7.8 of Hayashi (2000) we guarantee the consistency and the asymptotic normality of the MTD-Probit MMC model MLE.

Given the blocks of the $Q$ matrix - the matrices $P^{(ij)}$ - which can be estimated using the MLE, we estimate the $\hat{\eta}_{jk}$ coefficients, as well, through the MLE method, i.e.

$$\hat{\boldsymbol{\eta}}_j = argmax \log L\left(\tilde{\boldsymbol{\eta}}_j\right) \quad (4.5)$$

where

- $\boldsymbol{\eta}_j \equiv \begin{bmatrix} \eta_{j0} & \eta_{j1} & \cdots & \eta_{js} \end{bmatrix}'$ is the $s+1$ dimensional vector of the true unknown parameters,

- $\tilde{\boldsymbol{\eta}}_j \equiv \begin{bmatrix} \tilde{\eta}_{j0} & \tilde{\eta}_{j1} & \cdots & \tilde{\eta}_{js} \end{bmatrix}'$ is the $s+1$ dimensional vector of a hypothetical value of $\boldsymbol{\eta}_j$,

- $\hat{\boldsymbol{\eta}}_j \equiv \begin{bmatrix} \hat{\eta}_{j0} & \hat{\eta}_{j1} & \cdots & \hat{\eta}_{js} \end{bmatrix}'$ is the $s+1$ dimensional vector of the MTD-Probit estimates of $\boldsymbol{\eta}_j$

and

$$\log L\left(\tilde{\boldsymbol{\eta}}_j\right) = \sum_{i_o,i_1,\cdots,i_s} n_{i_1 i_2,\cdots,i_s i_0} \log\left\{P\left(S_{jt}=i_o\right|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s)^{\Phi}\right\} = \sum_{i_o,i_1,\cdots,i_s} n_{i_1 i_2,\cdots,i_s i_0} \log \left\{\frac{G\left[\tilde{\eta}_{j0}+\tilde{\eta}_{j1}P\left(S_{jt}=i_o\right|S_{1t-1}=i_1)+\cdots+\tilde{\eta}_{js}P\left(S_{jt}=i_o\right|S_{st-1}=i_s)\right]}{\sum_{k=1}^{m} G\left[\tilde{\eta}_{j0}+\tilde{\eta}_{j1}P\left(S_{jt}=k\right|S_{1t-1}=i_1)+\cdots+\tilde{\eta}_{js}P\left(S_{jt}=k\right|S_{st-1}=i_s)\right]}\right\} \quad (4.6)$$

It follows that we have

$$\sqrt{n}\left(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\right) \xrightarrow{d} N\left(0, \boldsymbol{\Sigma}_j\right) \quad (4.7)$$

where

- $\boldsymbol{\Sigma}_j = E\left[s\left(\boldsymbol{\eta}_j\right) s\left(\boldsymbol{\eta}_j\right)'\right]^{-1} = -E\left[\boldsymbol{\mathcal{H}}\left(\boldsymbol{\eta}_j\right)\right]^{-1}$, by information matrix equality.

- $\boldsymbol{\mathcal{H}}\left(\boldsymbol{\eta}_j\right) = \frac{\partial^2 \log L\left(\boldsymbol{\eta}_j\right)}{\partial \boldsymbol{\eta}_j \boldsymbol{\eta}_j'}$ is the Hessian matrix.

- $s\left(\boldsymbol{\eta}_j\right) = \frac{\partial \log L\left(\boldsymbol{\eta}_j\right)}{\partial \boldsymbol{\eta}_j}$ is the score vector.

Obviously the MTD-Probit estimator vector is consistent, i.e.

$$\hat{\boldsymbol{\eta}}_j \xrightarrow{p} \boldsymbol{\eta}_j \quad (4.8)$$

With regard to the inference problem, it is well known that there is a trinity of Chi-square distributed tests that can be applied in the context of MLE, namely the Wald test, the Lagrange multiplier test (LM) and the likelihood ratio test (LR), (Newey and McFadden 1993). Those tests can be used within the MMC framework to 1) test the joint significance of the parameters of the MMC, 2) test the independence between the categories and 3) determine the order of the MMC[5]. It may be interesting to evaluate,

[5] There is a lot of literature with regard to the determination of the order of the MC. For more details see Tong (1975), Katz (1981), Billingsley (1961), Bartllet (1952) and, more recently, Papapetrou and Kugiumtzis (2013), Csiszár et al (2000) and Zaho et al (2001). These results can also be easily extended to the MMC case.

for instance, the goodness of fit of the MMC. In particular one may test, among each equation, the global significance of the slope parameters vector $\boldsymbol{\eta}_j$, i.e.

$$H_o^j : \eta_{j1} = \eta_{j2} = \cdots = \eta_{js} = 0 \tag{4.9}$$

Under the null we have a test statistic following a $\chi^2_{(\alpha)}$ distribution with $\alpha = (n - s)$ degrees of freedom.

## 4.2 Multivariate Markov Chains: Forecast

In this section we briefly introduce the h-step-ahead MMC forecast problem. Since we have a homogeneous MMC the one-step-ahead forecast expression is quite simple. As soon,

$$P\left(S_{jt+1} = k|S_{1t}, \cdots, S_{st}\right) = P\left(S_{jt} = k|S_{1t-1}, \cdots, S_{st-1}\right) \tag{4.10}$$

Notwithstanding, the h-step-ahead MMC forecast, for $h > 1$, has a somewhat more troublesome expression. Using, firstly, the discrete case of Chapman-Kolmogorov equations and, secondly, the formula of total probability, for $h = 2$, it follows that:

$$\begin{aligned}
&P\left(S_{jt+2} = k|S_{1t}, \cdots, S_{st}\right)\\
&\quad= \sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m} P\left(S_{jt+2} = k, S_{1t+1} = i_1, \cdots, S_{st+1} = i_s|\, S_{1t}, \cdots, S_{st}\right)\\
&\quad= \sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m} P\left(S_{jt+2} = k|\, S_{1t+1} = i_1, \cdots, S_{st+1} = i_s, S_{1t}, \cdots, S_{st}\right)\\
&\qquad\times P\left(S_{1t+1} = i_1|S_{1t}, \cdots, S_{st}\right) P\left(S_{2t+1} = i_2|S_{1t+1} = i_1, S_{1t}, \cdots, S_{st}\right)\\
&\qquad\qquad\times \cdots \times P\left(S_{st+1} = i_s|S_{1t+1} = i_1, \cdots, S_{s-1t+1} = i_{s-1}, S_{1t}, \cdots, S_{st}\right)
\end{aligned} \tag{4.11}$$

for $h = 3$ we have:

$$\begin{aligned}
&P\left(S_{jt+3} = k|S_{1t}, \cdots, S_{st}\right)\\
&\quad= \sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m} P\left(S_{jt+3} = k, S_{1t+2} = i_1, \cdots, S_{st+2} = i_s|\, S_{1t}, \cdots, S_{st}\right)\\
&\quad= \sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m} P\left(S_{jt+3} = k|\, S_{1t+2} = i_1, \cdots, S_{st+2} = i_s, S_{1t}, \cdots, S_{st}\right)\\
&\qquad\times P\left(S_{1t+2} = i_1|S_{1t}, \cdots, S_{st}\right) P\left(S_{2t+2} = i_2|S_{1t+2} = i_1, S_{1t}, \cdots, S_{st}\right)\\
&\qquad\qquad\times \cdots \times P\left(S_{st+2} = i_s|S_{1t+2} = i_1 \cdots, S_{s-1t+2} = i_{s-1}, S_{1t}, \cdots, S_{st}\right)
\end{aligned} \tag{4.12}$$

The problem, however, lies in the circumstance that the probability terms regarding $S_{jt+l},\ l > 1$ are not $\mathcal{F}_t - measurable$ since the terms date explicitly to the present and to the future. Even if we assume a first order homogeneous MMC the expressions are unfeasible. Consequently, we need to put some assumptions in place in order to be able to manipulate the expressions (4.11) and (4.12). This is an issue that will be addressed in the next section.

# 5 Multivariate Markov Chains as Regressors: a new approach

*Model specification is an art as much as a science*
Russel Davidson and James G. Mackinnon
*Estimation and Inference in Econometrics*

## 5.1 Theoretical Model and Assumptions

Our model is based on the observation that if a MMC model is able to produce worthy forecasts of correlated data sequences then, if it is known that the MMC categories Granger cause a particular random process, for instance $y_t$, why not specify a functional form such as $y_t = m\left(S_{1t-l}, \cdots, S_{st-l}\right)$ in order to produce a worthy forecast of $y_{t+h}$?

Traditionally, and so far, the published literature only addresses the MMC as an end in itself. Here we propose a different and innovative concept: the usage of MMC as regressors in a certain model. Thus, given that the MMC Granger causes a specific dependent variable, and taking advantage of the information about the past state interactions between the MMC categories, we seek to forecast the current dependent variable more accurately.

Firstly, one must convert the categories $S_{jt}$ into a panoply of dummy variables. Given a first order MMC with $s$ categories and $m$ states we affect, for each category, a one state to one dummy variable, as follows:

$$z_{kjt} = 1\left\{S_{jt} = k\right\} \tag{5.1}$$

Where $1\{\cdot\}$ is the indicator function, $1\left\{S_{jt} = k\right\} = 1$ if $S_{jt} = k$ and $0$ otherwise.

Let us now assume, without any loss of generality, a linear specification like:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\delta} + u_t \tag{5.2}$$

where:

- $\boldsymbol{x}_t'$ may be a vector of both deterministic and stochastic components, like AR(1) or other $\mathcal{F}_{t-1}$ measurable predetermined terms.
- $\boldsymbol{z}_t'$ is a vector of dummy variables $z_{kjt}$ , concerning the MMC, defined in (5.1).

- $\{u_t\}$ is white noise process mean independent of $\boldsymbol{x}_t'$ and $\boldsymbol{z}_t'$. We do not assume any distribution for $u_t$.

Our goal is to forecast $y_{t+h}$ through its conditional mean, in other words we are interested in:

$$E\left(y_{t+h}\middle|\mathcal{F}_t\right) = E\left(\boldsymbol{x}_{t+h}'\middle|\mathcal{F}_t\right)\boldsymbol{\beta} + E\left(\boldsymbol{z}_{t+h}'\middle|\mathcal{F}_t\right)\boldsymbol{\delta} \tag{5.3}$$

given the exogeneity of the disturbance term, i.e. $E\left(u_t\middle|\mathcal{F}_t\right) = 0\ \forall t$

Unwinding the vector $\boldsymbol{z}_t'$ and the vector $\boldsymbol{\delta}$ it follows that:

$$\begin{aligned} y_{t+h} = \boldsymbol{x}_{t+h}'\boldsymbol{\beta} + \delta_{11}1\{S_{1t+h}=1\} + \cdots + \delta_{1m-1}1\{S_{1t+h}=m-1\} + \\ \delta_{21}1\{S_{2t+h}=1\} + \cdots + \delta_{2m-1}1\{S_{2t+h}=m-1\} + \cdots + \\ \delta_{s1}1\{S_{st+h}=1\} + \cdots + \delta_{sm-1}1\{S_{st+h}=m-1\} + u_t \end{aligned} \tag{5.4}$$

$$\Leftrightarrow$$

$$\begin{aligned} y_{t+h} = \boldsymbol{x}_{t+h}'\boldsymbol{\beta} + \delta_{11}z_{11t} + \cdots + \delta_{1m-1}z_{1m-1t} + \\ \delta_{21}z_{21t} + \cdots + \delta_{2m-1}z_{2m-1t} + \cdots + \delta_{s1}z_{s1t} + \cdots + \delta_{sm-1}z_{sm-1t} + u_t \end{aligned} \tag{5.5}$$

where $S_{jt}$ represents the $j-th$ categorical series of the MMC.

Nonetheless, with regard to the h-step-ahead forecast formulas, since, as we saw in the previous section, on the right-hand side of conditional probabilities, there are conditioning terms that are not $\mathcal{F}_t - measurable$ . Hence, to make the h-step-ahead terms feasible, we need to assume the following hypotheses:

**Assumption 5.1.** *First order MMC*

Let us assume that, given the $\sigma - algebra$ generated by the available information until $t-1$ - $\mathcal{F}_{t-1}$, we have

$$P\left(S_{jt}=k\middle|\mathcal{F}_{t-1}\right) = P\left(S_{jt}=k\middle|S_{1t-1}=i_1,\cdots,S_{st-1}=i_s\right) \tag{5.6}$$

That is, given the entire history of the multivariate stochastic process:

$$\{(S_{1t},\cdots,S_{st}),\ t=1,\cdots,T\} \tag{5.7}$$

only the immediate past is relevant both to explain and to forecast $S_{jt+1}$.

**Assumption 5.2.** *Homogeneous MMC*

Suppose that, under the system stability, we have a homogeneous MMC in the sense that

$$P\left(S_{jt}=k\middle|S_{1t-1},\cdots,S_{st-1}\right) = P\left(S_{jt+h}=k\middle|S_{1t+h-1},\cdots,S_{st+h-1}\right) \tag{5.8}$$

**Assumption 5.3.** *Ergodic MMC*

Let us assume an ergodic MMC, as stated in Theorem (4.2)

**Assumption 5.4.** *Contemporaneously negligible terms*

Consider that, given the past, the present values of the MMC are contemporaneous independent. Mathematically:

$$P\left(S_{jt}=k|\,S_{1t}=i_1,\cdots,S_{1s}=i_s,S_{1t-1},\cdots,S_{st-1}\right)=P\left(S_{jt}=k|\,S_{1t-1},\cdots,S_{st-1}\right) \tag{5.9}$$

So, in (4.11) and in (4.12) the conditioning terms $S_{jt+l},\ l>0$ vanish, given the terms $S_{jt}$, for being contemporaneous.

Under Assumptions 5.1 to 5.4 we have:

$$\begin{aligned}
&P\left(S_{jt+2}=k|S_{1t},\cdots,S_{st}\right)\\
&\quad=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+2}=k|\,S_{1t+1}=i_1,\cdots,S_{st+1}=i_s,\ \underbrace{S_{1t},\cdots,S_{st}}_{negligible\,(Ass\,5.1)}\right)\\
&\quad\times P\left(S_{1t+1}=i_1|S_{1t},\cdots,S_{st}\right)P\left(S_{2t+1}=i_2|\ \underbrace{S_{1t+1}=i_1}_{negligible\,(Ass\,5.4)}\ ,S_{1t},\cdots,S_{st}\right)\\
&\quad\times\cdots\times P\left(S_{st+1}=i_s|\underbrace{S_{1t+1}=i_1\cdots,S_{s-1t+1}=i_{s-1}}_{negligible\,(Ass\,5.4)},S_{1t},\cdots,S_{st}\right)\\
&\quad=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+2}=k|\,S_{1t+1}=i_1,\cdots,S_{st+1}=i_s\right)\\
&\quad\times P\left(S_{1t+1}=i_1|S_{1t},\cdots,S_{st}\right)P\left(S_{2t+1}=i_2|S_{1t},\cdots,S_{st}\right)\\
&\quad\times\cdots\times P\left(S_{st+1}=i_s|S_{1t},\cdots,S_{st}\right)
\end{aligned} \tag{5.10}$$

Similarly,

$$
\begin{aligned}
&P\left(S_{jt+3}=k|S_{1t},\cdots,S_{st}\right)\\
&=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+3}=k|\,S_{1t+2}=i_1,\cdots,S_{st+2}=i_s,\ \underbrace{S_{1t},\cdots,S_{st}}_{negligible\,(Ass\,5.1)}\right)\\
&\times P\left(S_{1t+2}=i_1|S_{1t},\cdots,S_{st}\right)P\left(S_{2t+2}=i_2|\ \underbrace{S_{1t+2}=i_1}_{negligible\,(Ass\,5.4)}\ ,S_{1t},\cdots,S_{st}\right)\\
&\times\cdots\times P\left(S_{st+2}=i_s|\underbrace{S_{1t+2}=i_1\cdots,S_{s-1t+2}=i_{s-1}}_{negligible\,(Ass\,5.4)},S_{1t},\cdots,S_{st}\right)\\
&=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+3}=k|\,S_{1t+2}=i_1,\cdots,S_{st+2}=i_s\right)\\
&\times\underbrace{P\left(S_{1t+2}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,(5.10)}\underbrace{P\left(S_{2t+2}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,(5.10)}\\
&\times\cdots\times\underbrace{P\left(S_{2t+2}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,(5.10)}
\end{aligned}
\tag{5.11}
$$

For a generic $h>3$,

$$
\begin{aligned}
&P\left(S_{jt+h}=k|S_{1t},\cdots,S_{st}\right)\\
&=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+h}=k|\,S_{1t+h-1}=i_1,\cdots,S_{st+h-1}=i_s,S_{1t},\cdots,S_{st}\right)\\
&=\sum_{i_1}^{m}\sum_{i_2}^{m}\cdots\sum_{i_s}^{m}P\left(S_{jt+h}=k|\,S_{1t+h-1}=i_1,\cdots,S_{st+h-1}=i_s\right)\\
&\times\underbrace{P\left(S_{1t+h-1}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,h-1}\underbrace{P\left(S_{2t+h-1}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,h-1}\\
&\times\cdots\times\underbrace{P\left(S_{st+h-1}=i_1|S_{1t},\cdots,S_{st}\right)}_{from\,h-1}
\end{aligned}
\tag{5.12}
$$

Even expression (5.12) only involves $\mathcal{F}_{t-1}$ measurable components. It should be noted that while Assumption 5.3 guarantees that the last expression exists, Assumption 5.2 guarantees that its components are known. In essence, here we have developed a simple strategy to get rid of the disturbing terms, i.e. terms that kept us from moving forward to achieve our goals: the closed form of h-step ahead forecast expressions. As the process is recursive, we just need to successively dovetail the expressions above to obtain expression (5.12).

## 5.2 Monte Carlo Simulation Study

### 5.2.1 Monte Carlo Simulation Study: Procedure and Design

In this section we evaluate the MMC predictive potential through a Monte Carlo simulation problem. The goal is to construct a model where the MMC, transformed into $s \times (m-1)$ dummy variables (one dummy for each state minus one, for each category), play the role of covariates, seeking to gauge how they help forecast a certain dependent variable. That is, what if we conceptualize regressors as a MMC? Will we achieve good results in terms of forecasts?

We consider here a simple process with two categories ($s = 2$) with each one taking values on 1, 2 or 3 ($m = 3$). We simulate the MMC, using the GAUSS program, in accordance with the following algorithm:

1. Initialize the process $\{(S_{1t}, S_{2t})\}$ by assigning arbitrary values for $S_{10}$ and for $S_{20}$
2. Simulate a continuous random variable that is uniformly distributed - $W \sim U(0,1)$
3. Define the $m^s \times m$ TPM whose elements are the probabilities
$$P\left(S_{1t} = i_o \middle| S_{1t-1} = i_1, S_{2t-1} = i_2\right) \tag{5.13}$$
(see the definition of the data-generating process below)
4. Given the initial values $S_{10}$ and $S_{20}$ (step 1), simulate the multivariate process $\{(S_{1t}, S_{2t})\}, t = 1, .., T$ as follows:
   (a) Let us define $p_i \equiv P\left(S_{1t} = i \middle| S_{1t-1} = i_1, S_{2t-1} = i_2\right)$
   (b) $S_{1t} = \begin{cases} 1 & \text{if} \quad 0 \le W < p_1 \\ 2 & \text{if} \quad p_1 \le W < p_1 + p_2 \\ 3 & \text{if} \quad p_1 + p_2 \le W < 1 \end{cases}$
   (c) Generate $S_{2t} \sim DU(1,3)$ $\big(DU(a,b)$ represents the discrete uniform distribution defined between $a$ and $b\big)$. Since our main focus in on the process$\{(S_{1t})\}$, we can generate $\{(S_{2t})\}$ from a simple probabilistic structure.
5. Repeat the steps 1-4 until $t = T$.

Thus, we construct our 4 dummy variables, as in (5.1), such that: $z_{jk,t} = 1\{S_{jt} = k\}, k = 1, \cdots, m-1$.

We consider the following linear data-generating process (DGP)

$$y_t = \boldsymbol{x}_t' \boldsymbol{\mu} + \boldsymbol{z}_t' \boldsymbol{\delta} + u_t \tag{5.14}$$

where

- $\boldsymbol{z}_t' \equiv \begin{bmatrix} z_{11} & z_{12} & z_{21} & z_{22} \end{bmatrix}$

- $\boldsymbol{\delta} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}'$, for simplicity

- $\boldsymbol{x}_t' = \begin{bmatrix} 1 & x_t \end{bmatrix}$ and $x_t | \mathcal{F}_{t-1} \sim N(0,1)$

- $\boldsymbol{\mu} = \begin{bmatrix} 1 & 1 \end{bmatrix}'$

- $u_t | \mathcal{F}_{t-1} \sim N(0,1)$

To fully define the DGP, we arbitrarily construct the TMP:

$$\begin{bmatrix} P(S_{1t}=1|S_{1t-1}=1,S_{2t-1}=1) & \cdots & P(S_{1t}=3|S_{1t-1}=1,S_{2t-1}=1) \\ P(S_{1t}=1|S_{1t-1}=1,S_{2t-1}=2) & \cdots & P(S_{1t}=3|S_{1t-1}=1,S_{2t-1}=2) \\ \vdots & \ddots & \vdots \\ P(S_{1t}=1|S_{1t-1}=3,S_{2t-1}=2) & \cdots & P(S_{1t}=3|S_{1t-1}=3,S_{2t-1}=2) \\ P(S_{1t}=1|S_{1t-1}=3,S_{2t-1}=3) & \cdots & P(S_{1t}=3|S_{1t-1}=3,S_{2t-1}=3) \end{bmatrix} \tag{5.15}$$

as follows

$$\begin{bmatrix} 0.85 & 0.10 & 0.05 \\ 0.10 & 0.85 & 0.05 \\ 0.60 & 0.20 & 0.20 \\ 0.80 & 0.10 & 0.10 \\ 0.10 & 0.80 & 0.10 \\ 0.10 & 0.20 & 0.70 \\ 0.70 & 0.15 & 0.15 \\ 0.10 & 0.70 & 0.20 \\ 0.10 & 0.10 & 0.80 \end{bmatrix} \tag{5.16}$$

We aim to compare the dependent variable $h$-step-ahead forecast errors produced by three different models:

1. Assuming that the values of dummy variables at $t+h$ are known,

$$\hat{z}_{jkt+h}^{(1)} = z_{jkt+h} \tag{5.17}$$

2. Assuming that the values of dummy variables at $t+h$ is its marginal mean (adjusted to an integer number)

$$\hat{z}_{jkt+h}^{(2)} = 1\left\{T^{-1}\sum_{t=1}^{T} z_{jkt} \geq Max_k\{\vartheta_{jk}\}\right\} \tag{5.18}$$

3. Assuming that the MMC is a homogeneous stochastic process and forecasting the value of dummy variables at $t+h$ such that:

$$\hat{z}_{jkt+h}^{(3)} = 1\left\{P\left(S_{jt+h} = k \middle| S_{1t} = i_1, S_{2t} = i_2\right)^{\hat{\Phi}} \geq Max_k\left\{\varsigma_{jk}\right\}\right\} \tag{5.19}$$

where

- $Max_k\{\vartheta_{jk}\} \equiv Max_k\left\{T^{-1}\sum_{t=1}^{T} z_{jkt}\right\}$
- $Max_k\{\varsigma_{jk}\} \equiv Max_k\left\{P\left(S_{jt+h} = k \middle| S_{1t} = i_1, S_{2t} = i_2\right)^{\hat{\Phi}}\right\}$
- $P\hat{(\cdot)}^{\Phi}$ is the MTD-Probit estimator of $P(\cdot)$.

and, as discussed in section 4, the right-hand side of equation (5.19) was estimated as follows:

$$P\left(S_{jt+1} = i_o \middle| S_{1t} = i_1, S_{2t} = i_2\right) = \frac{\Phi\left[\eta_{j0} + \eta_{j1}P\left(S_{jt+1} = i_o \middle| S_{1t} = i_1\right) + \eta_{j2}P\left(S_{jt+1} = i_o \middle| S_{2t} = i_2\right)\right]}{\sum_{k=1}^{3}\Phi\left[\eta_{j0} + \eta_{j1}P\left(S_{jt+1} = k \middle| S_{1t} = i_1\right) + \eta_{j2}P\left(S_{2t+1} = k \middle| S_{2t} = i_2\right)\right]} \tag{5.20}$$

As far as the MMC estimation is concerned, first of all we need to estimate the $P^{(ik)}$ matrices, whose generic elements are the quantities:

$$P\left(S_{jt+1} = i_o \middle| S_{kt} = i_1\right) \text{ for } j,k = 1,\cdots,s \text{ and } i_o, i_1 = 1,\cdots,m. \tag{5.21}$$

As stated before, those matrices are estimated consistently using the MLE method (expression 3.15).

Thereafter, we can consistently estimate the parameters $\hat{\eta}_{jk}$, in the second step, through the MTD-Probit model, using the MLE method.

Lastly, we compare the three forecast error measures produced by the different models. We consider two forecast error measures for each model:

1. The $MAE_{lh} = N^{-1}\sum\left|\hat{e}_{nlh}\right|$ and
2. The $RMSE_{lh} = \left(N^{-1}\sum\hat{e}_{nlh}^2\right)^{\frac{1}{2}}$

where $N$ is the number of replications considered in the experiment and $e_{nlh}$ is the $n-th$ replication forecast error produced by model $l$ $(l = 1,2,3)$ at the $h-th$ forecast step, i.e.

$$\hat{e}_{nlh} \equiv y_{t+hn} - \hat{y}_{t+hn}^{(l)} \tag{5.22}$$

where

$$\hat{y}_{t+hn}^{(l)} \equiv \boldsymbol{x}_{t+h}'\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{z}}_{t+h}^{(l)'}\hat{\boldsymbol{\delta}} \tag{5.23}$$

and

$$\hat{\boldsymbol{z}}_{t+h}^{(l)'} \equiv \left[\begin{array}{cccc} z_{11t+h}^{(l)} & z_{12t+h}^{(l)} & z_{21t+h}^{(l)} & z_{22t+h}^{(l)} \end{array}\right], \text{ for } l = 1,2,3. \tag{5.24}$$

### 5.2.2 Monte Carlo Simulation Study: Discussion of Results

In this section we report the results of the Monte Carlo study presented in the previous section, which investigates the potential forecast gains of a dependent variable, derived by processing categorical interrelated regressors as a MMC, i.e. by exploiting intra and inter-transition probabilities between categorical regressors. Theoretically, model 1, which assumes that the future values of dummy variables are known, leads to the best results in terms of forecasts, therefore we take the following ratios of forecast error measures:

$$\theta_p \equiv \frac{MAE_p}{MAE_1} \text{ and } \eta_p \equiv \frac{RMSE_p}{RMSE_1} \tag{5.25}$$

Implementing these specifications the error measure functions have a relative interpretation: a ratio of, for instance, $\eta_3 = 1.02$ implies that the $RMSE$ of the MMC is 2% worse than if we assume that we know the future (true values). Obviously we have $p = 2, 3$ since model 1 is our benchmark model. Table (5.1) presents the results of the forecast errors.

**Table 5.1:** Monte Carlo Experiment: Results of the Forecast Errors

| $T$ | Marginal mean | MTD-Probit $h=1$ | $h=2$ | $h=3$ | $h=4$ |
|---|---|---|---|---|---|
| | $\eta_3$ | $\eta_2$ | $\eta_2$ | $\eta_2$ | $\eta_2$ |
| 30 | 1.58973 | 1.11271 | 1.11271 | 1.11271 | 1.11271 |
| 500 | 1.42446 | 1.00000 | 1.00006 | 1.00993 | 1.00993 |
| 5000 | 1.40427 | 1.00042 | 1.00104 | 1.00067 | 1.00065 |
| | $\theta_3$ | $\theta_2$ | $\theta_2$ | $\theta_2$ | $\theta_2$ |
| 30 | 1.61977 | 1.10843 | 1.10843 | 1.10843 | 1.10843 |
| 500 | 1.48135 | 1.00000 | 1.00005 | 1.00947 | 1.00947 |
| 5000 | 1.44068 | 1.00087 | 1.00122 | 1.00098 | 1.00096 |

$T$ indicates the the sample size.

$h$ represents the h-step-ahead forecast considered.

We considered one-step-ahead forecasts for the marginal mean model.

We investigate both the finite and large sample performance of the proposed model in a Monte Carlo simulation problem. We considered sample sizes with $T = 30$, 500 and 5000 and, for each one, we performed 5000 Monte Carlo realizations, or replications.

For small samples ($T = 30$) with 5000 Monte Carlo realizations the MTD-Probit model results are roughly speaking 11% worse than the true values, while the marginal mean model (model 2) results are almost 60% worse. This implies that the marginal mean results are, at least, 45% worse than the MTD-Probit model results - taking the ratios.

For large samples, the actual results achieved are noteworthy: the point forecasts produced by MTD-Probit model are much closer to the true values. Generally, the MTD-Probit model results are less than 1% worse than the true values. This implies that the marginal mean model outcomes are generally 40% worse than the MTD-Probit model results. It can be said that the MTD-Probit model performs quite well. This fact suggests that if we take advantage of the deep past interaction between variables, i.e. if we capture intra and inter-transition probabilities within and between data categories, thus modeling these variables as a MMC, then we are able to achieve excellent results in terms of forecasts.

## 5.3 From theory to the real world: an Economic Illustration

*If European economic policy makers, like medical doctors, had to swear "to do no harm", they would all be banned from "practicing" economics.*
Mark Blyth,
*Austerity: The history of a dangerous idea.*

We are living in unprecedented troublesome times. In fact, the euro-zone, particularly its peripheral countries, are in the grip of a serious economic, social, political and institutional crisis. This crisis scenario has been mainly generated by the incapacity of both national and supranational (i.e. European Union) institutions to solve the sovereign debt crisis and the consequent imbalance of the financial and economic arenas caused by it. Portugal, Greece and Ireland have received formal bailout plans, while Italy and Spain have received informal ones from the Troika (IMF, ECB and EU). These bailout plans, yielded upon the condition of the implementation of austerity packages, which imply huge cuts in social expenditure, enormous decreases in wages and pensions, big tax rises and redundancies, have been throwing a significant number of people into unemployment and generalized poverty. The implementation of austerity measures has been happening with a broad consensus - among most politicians, economists, media and bankers. The problem is that the evidence shows that not only has the sovereign debt not been lowered, but also the structural imbalances remain.

Credit rating agencies (CRAs) are believed to be independent institutions that evaluate the financial credit risk of products like governments and corporate bonds, stocks and collateralized securities through the assignment of an ordinal scale (Table A.1). This scale represents the risk associated to a certain product and there is an inverse relationship between the risk inherent to a product and the return required by investors in this product: a lower rating suggests a higher risk, consequently, higher returns are required. CRAs have played a crucial role in this context due to the fact that their negative outlooks and rating downgrades have had a strong negative impact on sovereign bond yields.

According to the IMF 2010 Global Financial Stability Report (GFSR), CRA have contributed to this financial instability. *Rating downgrades can lead to knock-on and spillover effects that destabilize financial markets* (IMF 2010:86). The Report recommends that policymakers should reduce government dependence on credit ratings as much as possible, stating that *policymakers should continue their efforts to reduce their own reliance on credit ratings* (IMF 2010:112).

It is widely known that Capital World Investors (CWI) is the major shareholder of Standard & Poor's (S&P), and is also one of the biggest shareholders of Moody's, and is also one of the biggest institutional holders of southern countries' sovereign debt. This fact constitutes a conflict of interest[6], as the CRA assesses the risk, while at the same time they benefit from the evolution of this risk. For instance, a CRA can sell bonds on the secondary market and subsequently may decrease the ratings, which will quickly guarantee abnormal returns and profits. Due to these observations, we expect that: 1) we will find some structure with regard to the yields market, i.e. some predictable pattern on yields, 2) ratings contain information regarding yields, in such way that the former might help to forecast the latter, i.e. that there is a causality relationship from Portuguese ratings to Portuguese sovereign bond yields, and 3) due to the high level of European market integration, there are spillover effects from another southern European countries' ratings to Portuguese bond yields, thus there is a causality relationship from these ratings to Portuguese yields.

The next section describes the data. On the one hand, it is clear that the ratings are positively autocorrelated: a downgrade in a rating is followed by a downgrade or by maintenance (only in one situation did an upgrade follow a downgrade). In other words, the ratings show a persistent pattern over time. On the other hand, the ratings show

[6] Another evident conflict of interest is due to the corporate bond market. When a company issues bonds, the inherent rating assessment is a compulsory procedure and a CRA is attributed responsibility for carrying out this assessment. If the CRA returns an unfavorable rating, then the corporation will choose another CRA for another assessment. The CRA evaluator has a tendency and a factual incentive to overrate the corporate bond.

an explicit pattern of co-movements, which suggest that they are correlated among one another. Consequently, it is plausible to assume that southern European countries' ratings follow a MMC process. For simplicity, and without any loss of generalization, we assume a first order MMC, but the model can be easily adapted to accommodate a high-order MMC (HOMMC).

In this section we illustrate the benefits of treating interrelated categorical regressors as a MMC through an economic illustration. This illustration has two main objectives: 1) to analyze the predictability of Portuguese sovereign bond yields, in relation to the S&P sovereign credit ratings, 2) to understand whether the results obtained through a Monte Carlo simulation study, in the previous section, are confirmed empirically.

Here we shall consider an empirical example involving the Portuguese sovereign bond yields as a dependent variable and the PSI 20 stock market returns and sovereign credit ratings of some peripheral European countries as regressors. More details about the data will be explained in the next section.

### 5.3.1 The Data

Daily data on Portuguese sovereign bond yields, Portuguese stock index and S&P sovereign credit ratings (namely on Portugal, Spain, Italy, Greece and Ireland) from 2000 to 2012 are used in the analysis. While the data for yields was provided by the Bank of Portugal, data for the PSI20 stock index was taken from Reuters. Regarding the credit ratings, up until 2010 we used the data from Afonso et al (2012) and after 2010 the data was obtained from S&P. Given that the ratings are expressed on a non-numeric scale, we converted the ratings into a numeric ordinal scale through a linear transformation, following Afonso (2012), as in table (A.1). We denote the numeric ordinal rating for the $j-th$ country, at the instant $t$, as $V_{jt}$, $j = 1, 2, 3, 4$ and $t = 1, \cdots, T$.

Table (A.2) describes the data, table (A.3) presents some basic descriptive features of the data. Figure (A.1) shows yields, both the original series and the first differentiated series, while Figure (A.2) displays the remaining data time series plots. It should be noted that the graphs suggest some co-moments between the data. In fact, during the recent crisis period, while the ratings, in general, seem to exhibit a downward trend, the yields have an opposite behavior since they present a pronounced upward trend and their first differences display an increase in volatility. Regarding the PSI20 stock index returns, as well as the yields' first differences, we observe an increase in volatility, albeit, it seems to anticipate the volatility of the yields' first differences. Moreover, all the series present a negative skewness, with the exception of the yields - for obvious reasons, which indicates that negative extreme values are more likely to occur than positive extreme values. Lastly, in general, the processes present quite an excessive kurtosis (regarding the standard normal distribution). This leptokurtic feature means that extreme values are generated with relatively high probability.

There are, however, some caveats regarding the scale used for the ratings. In particular, there are many states that are not visited by the process, if we assume 17 states. Indeed, we need to group the series into a scale with fewer states, for example, 4.

$$S_{jt} = \begin{cases} 1 & if \quad V_{jt} \leq q_{j1} \\ 2 & if \quad q_{j1} < V_{jt} \leq q_{j2} \\ 3 & if \quad q_{j2} < V_{jt} \leq q_{j3} \\ 4 & if \quad V_{jt} > q_3 \end{cases} \tag{5.26}$$

We choose the quantities $q_{ji}$ that lead us to the best results in terms of forecast and that assure a quite even distribution of the variables $S_{jt}$ . Table (A.4) displays the chosen $q_{ji}$ for estimation purposes. Later, we create four dummy variables for each country, where each dummy represents a state. Mathematically we have:

$$z_{jkt} = 1\{S_{jt} = k\}\ k = 1, \cdots, m; j = 1, \cdots s;\ t = 1, \cdots, T. \tag{5.27}$$

We consider 5 categorical series $(s = 5)$ and 4 states $(m = 4)$. It should be noted that we cannot estimate a a fully parametrized MMC in this context given the fact that it involves 4096 independent parameters $m^{s+1}$ and we just have 3140 observations.

A last observation concerns the erratic behavior of the yields. We fitted an $AR(1)$ on yields and we obtained an autoregressive coefficient of approximately $1$. Since all unit root tests (Table A.5) pointed out the non-stationary nature of yields we considered their first differences. It is known that we must look at DF tests with some caution. We must be particularly careful in choosing which determinist regressors to include in the ADF auxiliary regression. Regarding the choice of deterministic regressors, one should include a trend term when the process seems to have some kind of trend and, typically, it is the economic theory that dictates the inclusion, or not, of such deterministic regressors. Usually, concerning rates of all kinds, one should not include a trend in auxiliary regression. This is a problematic issue due to the tests possibly being unable to reject the null hypothesis of non-stationarity , even more so when yields seem to have a structural break (Banerjee et al, 1992; Stock, 1994). One should take into account that the power distortions of the DF test are further compounded by the presence of conditional heteroskedasticity, (Rodrigues and Rubia, 2005). Nevertheless, we performed the tests above but restricted the sample to 2008 and the results were consistent with the previous ones: the yields are an integrated process.

Nonetheless, given the exceptional economic conditions and the current uncertain environment, factors that contribute to the fact that apparently the yields exhibit an upward trend, we perform the tests with and without a linear deterministic trend.

### 5.3.2 Procedure and Design: Model Specification

Both yields and their first differences display ARCH type effects. So, we have fitted an ARMA model to each case and Table (A.6) shows this statement.

Moreover, yields seems to exhibit a positive trend in the recent period of high volatility. Consequently, the typical candidates to portray such behavior are the ARCH[7] (Engle, 1982) and the GARCH (Bollerslev, 1986) models. The primary idea is to treat the volatility as non-constant over time, thus we model here not only the random variable in mean but simultaneously in variance. We assume the following specification:

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\gamma} + \boldsymbol{z}_t'\boldsymbol{\delta} + u_t \\ u_t &= \sigma_t v_t \\ \sigma_t^2 &= \theta + \sum_{p=1}^{P} \alpha_p u_{t-p}^2 + \sum_{r=1}^{R} \beta_r \sigma_{t-r}^2 \end{aligned} \tag{5.28}$$

where:

- $\boldsymbol{x}_t'$ is a vector of both deterministic and stochastic components, like AR(1) or other predetermined $\mathcal{F}_{t-1} - measurable$ terms.
- $\boldsymbol{z}_t'$ is a vector of dummy variables
- $v_t|\,\mathcal{F}_{t-1} \sim t\,(v)$ with zero mean and unit variance

Due the some features of yields, where unconditional distribution seems to generate extreme events with relatively high probability (high kurtosis), we assume a $t - student$ distribution for the disturbance term $v_t$, which can accommodate this characteristic of the data, in order to carry out the estimation more efficiently.

Concerning our particular model specification we have selected a GARCH (1,1) and, of course, we do not need to include the whole dummy set as covariates. We have:

- $\boldsymbol{x}_t' \equiv \left[\begin{array}{cc} d\,(yields)_{t-1} & ret_{t-1} \end{array}\right]$,
- $\boldsymbol{z}_t' \equiv \left[\begin{array}{cccccc} z_{14t-1} & z_{24t-1} & z_{32t-1} & z_{34t-1} & z_{43t-1} & z_{53t-1} \end{array}\right]$ and
- $\alpha \equiv \alpha_1$ and $\beta \equiv \beta_1$

We should recall that, to understand the estimates better, we denote the countries as in table (5.2).

[7] ARCH type models are quite widespread and there are numerous published surveys on the models of the ARCH family, among which we recall, here, for instance, Bollerslev's (2010). A thorough presentation of the ARCH models goes beyond the scope of this thesis.

**Table 5.2:** Country Code

| Country | $j$ |
|---|---|
| Portugal | 1 |
| Italy | 2 |
| Spain | 3 |
| Greece | 4 |
| Ireland | 5 |

The first step consists of investigating whether the dummy variables help forecast yields, i.e. whether the yields are actually caused, Granger speaking, by peripheral European countries Standard and Poor's sovereign credit ratings, despite their coefficients remaining statistically significant. Hence, firstly, we compute Granger causality tests, and secondly we estimate two GARCH models (one with and another without rating dummy variables) and then we perform an out-of-sample forecast error analysis. The first GARCH model (model one), without dummy variables, is:

$$\begin{aligned} y_t &= \boldsymbol{x}_t' \boldsymbol{\gamma} + u_t \\ u_t &= \sigma_t v_t \\ \sigma_t^2 &= \theta + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 \end{aligned} \tag{5.29}$$

Equation (5.28) exhibits the second GARCH model (model 2).

Both tests confirm that the dummy Granger variables cause yields. In the first test (Table A.7) the hypothesis of non-causality is rejected for every significance level, both for yield levels and first differentiated yields. In the second test (Table A.8) there are significant forecast error gains, compared to the first differentiated yields, if we include the dummy variables in the model. In a nutshell, Standard and Poor's ratings contain valuable information concerning the forecasting process of the yields, so they should be included in the model.

In the second step, we address the issue of forecasting, as in a Monte Carlo simulation study. Accordingly, we need to fit a MMC in order to obtain the MMC one-step-ahead forecasts with regard to the dummy variables.

Table (5.3) summarizes the MMC estimations. Likewise in the Monte Carlo study section the estimations were carried out using the MDT-Probit method.

Finally, staying with the MMC estimation, regarding the goodness of the fit, it is important to note the global significance of the model. Indeed, through a Likelihood Ratio test, presented in section 3, we have evidence that the parameters are jointly significant (Table 5.4). The null hypothesis of $H_0^j : \eta_{j1} = \cdots = \eta_{j5} = 0$, $j = 1, \cdots, 4$, is rejected, in all equations, for any significance level.

**Table 5.3:** MMC Estimation

| Equation | $\hat{\eta}_{j0}$ | $\hat{\eta}_{j1}$ | $\hat{\eta}_{j2}$ | $\hat{\eta}_{j3}$ | $\hat{\eta}_{j4}$ | $\hat{\eta}_{j5}$ | *Mean LL* |
|---|---|---|---|---|---|---|---|
| 1 Rat-Pt | $-6.2524^{***}$ (0.2505) | $1.2144^{***}$ (0.1384) | $0.2858^{**}$ (0.1295) | $0.3785^{**}$ (0.1672) | $-0.1908$ (0.4116) | 0.1996 (0.2275) | $-0.0057$ |
| 2 Rat-It | $-6.2786^{***}$ (0.9223) | $0.7651^{*}$ (0.4818) | $1.4297^{***}$ (0.2803) | $-0.1244$ (0.3312) | $-1.1740^{*}$ (0.6669) | $1.0455^{***}$ (0.2146) | $-0.0054$ |
| 3 Rat-Sp | $-5.5640^{***}$ (1.0771) | 0.1621 (0.6099) | 2.2765 (1.8478) | $2.1667^{**}$ (1.1714) | 1.9078 (1.3290) | 0.3365 (0.5616) | $-0.0051$ |
| 4 Rat-Gr | $-4.8270$ (6.1509) | $-0.0056$ (0.0606) | $-0.5570$ (0.7015) | 0.0832 (0.8058) | 1.7814 (3.7020) | 1.0266 (1.8851) | $-0.0075$ |
| 5 Rat-Ir | $-5.9023^{***}$ (0.3379) | $0.8858^{*}$ (0.5062) | 0.0801 (0.4166) | $-0.4817$ (0.5383) | $-0.0383$ (0.1591) | $1.4070^{***}$ (0.1521) | $-0.0073$ |

Coefficient estimates are presented, standard errors between parentheses.

*Mean LL* represents the mean of the log-likelihood function.

$^{***}$,$^{**}$ and $^{*}$ indicates the statistical significance level, respectively, for 1%, 5% and 10%

**Table 5.4:** MMC Estimation: Goodness of Fit

| Equation | $LR - Statistic$ |
|---|---|
| 1 Rat-Pt | 6636.214 (0.000) |
| 2 Rat-It | 6637.704 (0.000) |
| 3 Rat-Sp | 6639.9102 (0.000) |
| 4 Rat-Gr | 6624.589 (0.000) |
| 5 Rat-Ir | 6626.261 (0.000) |

$LR^j_{obs}$ associated to $H^j_0 : \eta_{j1} = \cdots = \eta_{j5} = 0$ are reported.

$p - values$ in parentheses.

The next section briefly presents and discusses the remaining estimation and forecast results.

### 5.3.3 Discussion of Results

Table (5.5) displays the estimation results for the mean equation and for the variance equation. All coefficients are statistically significant apart from the dummy for Ireland. Moreover, the yields first differences presented are positively autocorrelated, since the autoregressive coefficient is positive and the PSI20 returns seem to have a negative impact on the yields' variation. This is an expected result: if the PSI20 returns decreased yesterday then we expect a positive variation on yields. With regard to the ratings' coefficient signs, if the Portuguese rating is in state 4 then we expect a future negative variation on the yields. With respect to Spain's coefficients, both state 2 and 4 have a positive impact on yields' future variations, but the magnitude of state 2 is considerably higher than that of the state 4, as expected. With regard to the variance equation, the $\alpha$ and $\beta$ estimates are also statistically significant, confirming the idea that yields' first differences exhibit both ARCH and GARCH effects. As we have $\hat{\alpha} + \hat{\beta} \approx 1$ we can say that, as the second order stationarity condition is not verified, the model estimated is not a second order stationary GARCH – it seems to be an IGARCH model. This is an expected result, given that, as it is known, a structural break in variance can lead to a spurious IGARCH model specification. Furthermore, unlike the generality of the ARMA family models, the GARCH models allow the possible coexistence of strict stationarity and non stationarity, speaking in second-order terms. Thus, it is straightforward to prove that the sufficient condition for strict stationarity

$$E\left[\log\left(\beta + \alpha v_t^2\right) < 0\right] \tag{5.30}$$

is verified by our model.

**Table 5.5:** GARCH Estimation Results

| Mean Equation | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma_1$ | $\gamma_2$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_4$ | $\delta_5$ | $\delta_6$ |
| $0.096^{***}$ (0.018) | $-0.235^{***}$ (0.069) | $-0.006^{***}$ (0.002) | $0.005^{*}$ (0.002) | $0.021^{***}$ (0.007) | $0.006^{***}$ (0.002) | $-0.004^{**}$ (0.002) | $-0.003$ (0.002) |
| Variance Equation | | | | | | | |
| | | $\theta$ | $\alpha$ | $\beta$ | $v$ | | |
| | | $2.70 \times 10^{-9***}$ ($6.52 \times 10^{-10}$) | $0.057^{***}$ (0.010) | $0.934^{***}$ (0.011) | $5.696^{***}$ (0.575) | | |

Coefficient estimates are presented, standard errors in parentheses.

$^{***}$,$^{**}$ and $^{*}$ indicates the statistical significance level, respectively, for 1%, 5% and 10%

With regard to the forecast errors, our economic illustration confirms the Monte Carlo Simulation study results. Thus, the idea of conceptualizing categorical regressors such as a MMC seems to work empirically. The MTD-Probit model forecasts are better than in the Monte Carlo study since the errors vanish. In fact, despite the out-of-sample size, the forecast errors of the MTD-Probit model are minimal - its fitted values equal the true values, i.e. the fitted values of model 1 (assuming that the future values of the dummy variables are known). A possible explanation for this result is due to the fact that the southern peripheral countries' ratings have changed little over the recent past. In any case, we have archieved excellent results. With regard to model 2 - marginal means - on average the forecast results are at least 2% worse than the results of model 3. Table (5.6) shows those results. It should be pointed out, as well, that with these specifications the heteroskedastic ARCH effects have been purged out of the model (Table A.9). The results obtained lead us to the following remarks:

**Table 5.6:** GARCH Estimation: Results of Forecast Errors

We constructed the forecasts in a sequential way. Firstly, $n$ dimensional out-of-sample are considered; Secondly, we produced successive $h$ step ahead forecasts until $t+n$ (the last out-of-sample observation is $t+n-h$).

| $n$ | Marginal Mean | | MTD-Probit | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $h=1$ | | $h=2$ | | $h=3$ | |
| | $\eta_3$ | $\theta_3$ | $\eta_2$ | $\theta_2$ | $\eta_2$ | $\theta_2$ | $\eta_2$ | $\theta_2$ |
| 30 | 1.0211 | 1.0138 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 50 | 1.0263 | 1.0293 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 100 | 1.0317 | 1.0338 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

$n$ indicates the out of sample dimension and $h$ represents the $h$ step ahead forecast considered.

1. We find a predictable pattern for Portuguese sovereign bond yields, therefore this specific market does not seem to be an efficient market as it appears to reveal some structure. Moreover, there might be abnormal excess return gains regarding this market. With regard to efficient bond market hypothesis (EBMH), which has become a well known and much discussed topic over recent decades, there is no consensus on the validity of EBMH. On the one hand are its advocates, of whom we highlight Fama (1975), Pesanto (1978), Kroon (1991) and Malkiel (2003). On the other hand, some critics such as Mishkin (1978), Zunino et al (2003), Shleifer (2003) and Bai et al (2013) have found some recent evidence of inefficiency in bond markets. The EMBH implies that the possession of past market information does not affect the future yields, i.e. that prices incorporate and instantaneously reflect all information (past public information, new information or even insider information, depending on the version)[8], consequently abnormal returns cannot occur. As noted by Mishkin, efficient-market theory implies that *returns in long-term bond and stock markets should be affected only by new information in the marketplace and should be uncorrelated with any past available information*, Mishkin (1978:712). Pesanto, in the same spirit, argues that *the bond market is efficient by testing its implication that forward rates pertaining to any*

[8] Traditionally, there are three variants of Efficient Market Hypothesis. The weak form states that the history path of a certain asset says nothing about its future path; the semi-strong form states that prices adjust instantaneously to significant news; the strong version asserts that even insider information, i.e. privileged information, is irrelevant to explain the future path of a certain asset.

*fixed date in the future follow a martingale sequence*, Pesanto (1978:1058). Figure (A.4) displays the correlogram of the yields. It is clear that both are correlated sequences. Moreover, we have found that past PSI20 returns contain information concerning future variations in yields: the yields are autocorrelated and are also negative correlated with past returns.

2. In addition, if one incorporates information regarding the S&P ratings on peripheral European countries, we are able to obtain forecasts that have a greater level of agreement. This is quite obvious, given the conflict of interests between CRAs and the yields market, due the fact that the major Stakeholder of CRAs - CWI - is, simultaneously, one of the biggest shareholders of peripheral European countries' sovereign debt, i.e. they produce, buy and sell recommendations on yields while at the same time they control the ratings which have a significant impact on the yields. Due to this point and to the previous one, peripheral countries' sovereign bond markets do not seem to be efficient markets.

3. Our main idea to conceptualize a MMC as covariates appears to be empirically feasible. Moreover, the results from the empirical illustration confirm those obtained by the Monte Carlo simulation study, showing a significant improvement on the results of the forecast errors.

# 6 Extensions and Further Research

With regard to further research, we present some ideas that could be of interest to investigate below.

- Extend the MMC methods, in general, to non-homogeneous MMC and, in particular, to elaborate on it in our framework - non-homgeneous MMC as covariates
- It could be interesting to investigate the situation in which the MMC arises as a result of continuous covariates, i.e. when $V_{jkt} \in \mathbb{R}$, despite the fact that the discretization of a given continuous variable implies loss of information. Namely, it would be appropriate to provide answers to the following question: are the benefits of the utilization of the MMC larger or lower than this loss of information?
- One may consider doing the opposite of what we performed in this thesis: instead of viewing a MMC as regressors, i.e. MMC as a function of a dependent variable, conceptualize a MMC that depends also on exogenous variables as follows:

$$P\left(S_{jt} = i_o \middle| \mathcal{F}_{t-1}\right) = P\left(S_{jt} = i_o \middle| S_{1t-1}, \cdots, S_{st-1}, \boldsymbol{x}'_{t-1}\right) \tag{6.1}$$

  where $\boldsymbol{x}'_{t-1}$ is a vector of $\mathcal{F}_{t-1} - measurable$ covariates.

# 7 Conclusions

This paper proposed a new concept: the usage of MMC as regressors. Obviously, our model is based on the assumption that the MMC Granger causes the respective dependent variable. Our main theoretical result is that the treatment of MMC as covariates works in the sense that it leads us to extremely good results in terms of forecasts. A Monte Carlo simulation study and empirical illustration support our theoretical result. With respect to the Monte Carlo simulation study, we conclude that, in large samples, on average, the MMC forecast errors are less than 1% worse than the true fitted values (assuming that the values of dummy variables at $t + h$ are known) and lead to forecast errors at least 40% better than the benchmark model (the marginal mean model). These outcomes are confirmed by our economic illustration since the MMC forecast errors of the dependent variable - Portuguese sovereign bond yields - equals its true fitted values. Furthermore, due to the empirical illustration we conclude that:

1. The policy of the CRAs has failed. The CRAs that rated the Lehman Brothers with triple A a few days before its bankruptcy, are the same CRAs that have

been contributing to the dramatic increase of peripheral countries' sovereign bond yields. These CRAs are supposed to be independent institutions. However, their political independence becomes merely apparent, given that they have a rooted conflict of interest. Due to this fact one obvious solution is to reject the utter power that these CRAs have nowadays. One useful way to question them and their influence is through the creation of a new CRA, drawn up under the jurisdiction of the European Union institutions and designed to meet the objectives of transparency and credibility necessary to guarantee that the financial systems of the countries that belong to the euro zone are solid and effective.

2. There is a consensus with regard to the payment of the peripheral countries' sovereign debt. The national governments, under pressure from the international institutions and the consequent bailout plans, are one of the decision centers that most contribute to this idea. Therefore, the large majority of the national governments have followed political guidelines based on crosscutting to enable the repayment of debts. This has led to governmental options concerning expenditure and revenue being constrained by this objective. However, and despite the fact that huge sacrifices are been made by most of the population, it should be noted, on the one hand, that the annual Portuguese debt service is larger than the expenditure on education or on health care. On the other hand, all the current economic policies regarding the debt payment have failed: the debt is still increasing, regardless of the social imbalances and the increasing phenomena of poverty and hunger. This should lead us to the following question: should the payment of the debt be the main political priority? We believe that this question cannot be answered without a prior discussion about the nature of the state and about the fundamental functions that a state must deliver.

# References


1. Adke, S. and Deshmukh, S. (1988). *Limit distribution of a high-order Markov Chain*. Journal of the Royal Statistical Society: Series B 50, 105–108.

2. Afonso, A., Furceri, D., Gomes, P. (2012). *Sovereign credit ratings and financial markets linkages: application to European data*, Journal of International Money and Finance 31 (3), 606-638.

3. Azzalini, A. (1994). *Logistic regression for autocorrelated data with application to repeated measures*. Biometrika 81 (4), 767-775.

4. Bartlet, M. (1952). *A Sampling test of the $\chi^2$ theory for probability chains*. Biometrika 39, 118-121.

5. Bai, J., Fleming, M. and Horan, C. (2013). *The Microstructure of China's Government Bond Market*. Federal Reserve Bank of New York Staff, Report No. 622.

6. Banerjee, A., Lumsdaine, R., and Stock, J. (1992). *Recursive and sequential tests of the unit-root and trend-break hypotheses: theory and international evidence*. Journal of Business & Economic Statistics 10(3), 271-287.

7. Berchtold, A. (1995). *Autoregressive modelling of Markov chains*. In Proc. 10th International Workshop on Statistical Modelling 19–26. Springer, New York.

8. Berchtold, A. (1996). *Modélisation autorégressive des chaînes de Markov: Utilisation d'une matrice différente pour chaque retard*. Rev. Statist. Appl. 44 5–25.

9. Berchtold, A. (1998). *Chaînes de Markov et Modèles de Transition: Applications aux Sciences Sociales*. HERMES, Paris.

10. Berchtold, A. (2001). *Estimation in the mixture transition distribution model*. Journal of Time Series Analysis 22, 379-397.

11. Berchtold, A., Raftery, A. (2002). *The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series*. Statistical Science 17, 328-356.

12. Billingsley, P. (1960). *Statistical Methods in Markov Chains*. In Stanford Meetings of Institute of Mathematical Statistics.

13. Billingsley, P. (1961). *Statistical Inference for Markov Processes.* Univ. Chicago Press.

14. Bollerslev, T. (1986). *Generalized Autoregressive Conditional Heteroskedasticity.* Journal of Econometrics 31, 307-327.

15. Bollerslev, T. (2008). *Glossary to ARCH (GARCH).* Technical report, Glossary to ARCH (GARCH). Research Paper 2008-49.

16. Chen, D. and Lio, Y. (2009). *A Novel Estimation Approach for Mixture Transition Distribution Model in High-Order Markov Chains. Communications* in Statistics - Simulation and Computation 38, 990 -1003.

17. Ching, W., Fung, E. and Ng, M. (2002). *A Multivariate Markov Chain Model for Categorical Data Sequences and its Applications in Demand Predictions.* Journal of Management Mathematics 13, 187-199

18. Ching, W. Zhang, S. and Ng, M.. (2007a). *On Multi-dimensional Markov Chain Models.* Pacific Journal of Optimization 3, 235-243.

19. Ching, W., Li, L., Li, T., and Zhang, S. (2007b). *A New Multivariate Markov Chain Model with Applications to Sales Demand Forecasting*: Proceedings of the 2007 International Conference on Industrial Engineering and Systems Management, Beijing.

20. Ching, W., Siu, T., Li, L., Jiang, H., Li, T. and Li., (2009). *An improved parsimonious multivariate Markov chain model for credit risk.* Journal of Credit Risk 5, 1–25.

21. Csiszár, I. and Shields, P. (2000). *The consistency of the BIC Markov order estimator*, The Annals of Statistics 28, 1601–1619.

22. Davidson, R. and J. MacKinnon (1993). *Estimation and Inference in Econometrics.* New York, Oxford University Press, 1993

23. Engle, R. (1982). *Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation.* Econometrica 50, 987-1008.

24. Fama, E., (1975). *Short-Term Interest Rates as Predictors of Inflation*, American Economic Review 65, 269-282.

25. Fama, E., (1976). *Inflation Uncertainty and Expected Returns on Treasury Bills*, Chicago Journals 84(3), 427-448.

26. Flournoy, N. (1990). *Dependency in multivariate markov chains.* Linear Algebra and its Applications 127, 85-106.

27. Fung, E., Ching, W., Chu, S., Ng, M. and Zang, W. (2002). *Multivariate Markov chain models.* 2002 IEEE International Conference on Systems, Man and Cybernetics Oct. 2002, 3 (5)

28. Gottschau, A. (1992). *Exchangeability in Multivariate Markov Chain Models.* Biometrics 48 (3), 751-763.

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

30. International Monetary Fund (2010). *Global Financial Stability Report: Sovereigns, Funding, and Systemic Liquidity.* IMF, November 2010.

31. Jacobs, P. and Lewis, P. (1978). *Discrete time series generated by mixtures. I. Correlational and runs properties.* Journal of the Royal Statistical Society: Series B 40, 94–105.

32. Katz, R. (1981). *On some criteria for estimating the order of a Markov chain*, Technometrics 23 (3) 243–249.

33. Kroon, E. (1991). *Bond market efficiency: some Dutch evidence.* Serie Research Memoranda from VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics no 45.

34. Kijima, M., Komoribayashi, K., Suzuki, E. (2002). *A Multivariate Markov Model for Simulating Correlated Defaults.* Journal of Risk 4(4), 1-32.

35. Kvam, P. and Sokol, J. S. (2006). *A logistic regression/Markov chain model for NCAA basketball.* Naval Research Logistics, 53, 788–803.

36. Le, N., Martin, R. and Raftery, A. (1996). *Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models.* J. Amer. Statist. Assoc. 91 1504–1515.

37. Lebre, S. and Bourguignon, P. (2008). *An EM algorithm for estimation in the mixture transition distribution model.* Journal of Statistical Computation and Simulation 78, 713-329.

38. Liu, T. (2010). *Application of Markov Chains to Analize and Predict the Time Series.* Modern Applied Science. 4 (5).

39. Logan, J. (1981). *A structural model of the higher-order Markov process incorporating reversion effects.* The Journal of Mathematical Sociology 8, 75–89.

40. Malkiel, B. (2003). The Efficient Market Hypothesis and Its Critics. The Journal of Economic Perspectives 17(1), 59-82.

41. Mehran, F. (1989). *Analysis of discrete longitudinal data: Infinite-lag Markov models. In Statistical Data Analysis and Inference* (Y. Dodge, ed.) 533–541. North-Holland, Amsterdam.

42. Mishkin, F. (1978). *Efficient-Markets Theory: Implications for Monetary Policy.* Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution 9(3), 707-752.

43. Newey W and McFadden D. (1994). *Large sample estimation and hypothesis testing.* In Handbook of Econometrics, Vol. IV, Engle R, McFadden D (eds). North Holland: Amsterdam, 2111-2245.

44. Nicolau, J. (2012). *Estimation and Inference in Multivariate Markov Chains.* (Submitted)

45. Nicolau, J. (2013). *A Method to Estimate Multivariate Markov Chains.* (Submitted)

46. Oz, E. and Erpolat, S. (2011). *An Application of Multivariate Markov Chain Model on the Changes in Exchanges Rates: Turkey Case.* European Journal of Social Sciences 18(4), 542-552.

47. Papapetrou, M. and Kugiumtzis, D. (2013). *Markov chain order estimation with conditional mutual information.* Physica A 392, 1593–1601.

48. Pegram, G. (1980). *An autoregressive model for multilag Markov chains.* Journal of Applied Probability 17, 350–362.

49. Pesanto (1978). *On the Efficiency of the Bond Market: Some Canadian Evidence.* Journal of Political Economy 86(6), 1057-1076.

50. Raftery, A. and Tavaré, S. (1994). *Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model.* Applied Statististics 43, 179–199.

51. Raftery, A. (1985). *A model for high-order Markov chains.* Journal of the Royal Statistical Society - Series B 47, 528-539.

52. Rajarshin, M. (2013). *Statistical Inference for Discrete Time Stochastic Processes.* Springer Briefs on Statistics. Springer.

53. Rodrigues, P. and Rubia, A (2005). *The performance of unit root tests under level-dependent heteroskedasticity.* Economics Letters 89(3), 262-268.

54. Shleifer (2003). *Will the Sovereign Debt Market Survive?.* American Economic Review 93(2), 85-90.

55. Siu, T., Ching, W., Fung, S. and Ng, M. (2005). *On a multivariate Markov chain model for credit risk measurement.* Quantitative Finance 5(6), 543-556.

56. Stock, J. H. (1994). *Unit roots, structural breaks and trends.* Handbook of econometrics 4, 2739-2841.

57. Tong, H. (1975). *Determination of the order of a Markov chain by Akaike's information criterion*, Journal of Applied Probability 12 (3), 488–497.

58. Wang, C., Huang, T. (2013). *A new improved parsimonious Multivariate Markov Chain Model.* Journal of Applied Mathematics 3, 1-17.

59. Wong, C. and Li, W. (2000). *On a mixture autoregression model.* J. Roy. Statist. Soc. Ser. B 62 95–115

60. Wasserman, S. and Pattinson, P. (1996). Logit models and logistic regressions for social networks: I. an Introduction to Markov Graphs and $p^*$. Psychometrika 61(3), 401-425.

61. Zhao, L., Dorea, C. and Gonçalves, C. (2001). *On determination of the order of a Markov chain*, Statistical Inference of Stochastic Processes 4, 273–282.

62. Zhang, S., Ching, W. and Ng, M. (2006). *On higher-dimensional markov chain models.* Pacific Journal of Optimization.

63. Zhang, S., Ching, W., Jiao, Y., Wu, L. and Chan, R. (2008). *Construction and Control of Genetic Regulatory Networks: A Multivariate Markov Chain Approach*, Journal of Biomedical Science and Engineering 1, 15-21.

64. Zhu, D. and Ching, W. (2010). *Multivariate Markov chain models for production planning.* International Journal of Intelligent Engineering Informatics 1(2), 156-173.

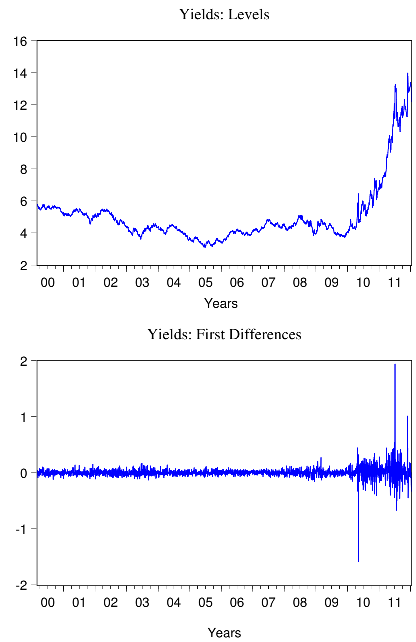
65. Zhu, D., Ching, W. (2010). *A New Estimation Method for Multivariate Markov Chain Model with Application in Demand Predictions.* In BIFE 10 Proceedings of the 2010 Third International Conference on Business Intelligence and Financial Engineering (eds L. Yu, K. Lai, S. Wang ), Hong Kong.

66. Zunino, L., Bariviera A., Guercio, M., Martinez, L. and Rosso, O. (2003), *On the efficiency of sovereign bond markets.* Physica A 391, 4342–4349.

# A Appendix

## A.1 Figures

## Figure A.1: Yields




## Figure A.2: Ratings and PSI20 Returns

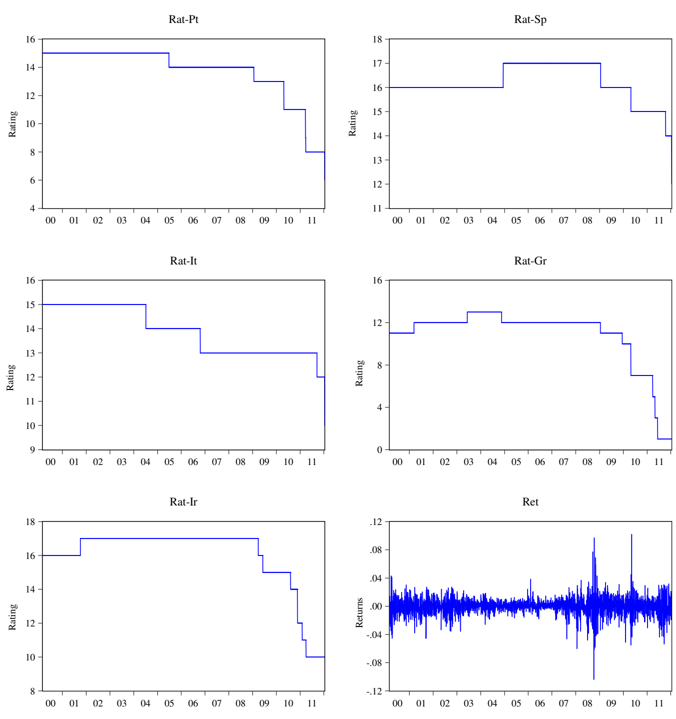


**Figure A.3:** GARCH Residuals

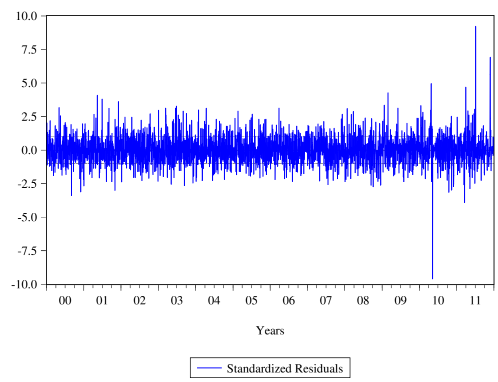
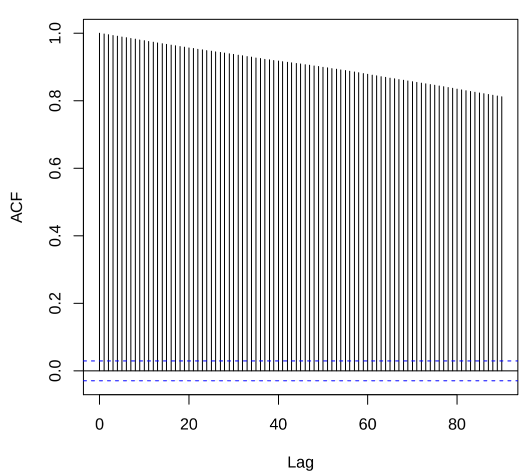



**Figure A.4:** Yields Correlogram



## A.2 Tables

**Table A.1:** Standard and Poor's Ratings Scale

| S&P Ratings | Numeric Scale |
|---|---|
| *AAA* | 17 |
| *AA+* | 16 |
| *AA* | 15 |
| *AA−* | 14 |
| *A+* | 13 |
| *A* | 12 |
| *A−* | 11 |
| *BBB+* | 10 |
| *BBB* | 9 |
| *BBB−* | 8 |
| *BB+* | 7 |
| *BB* | 6 |
| *BB−* | 5 |
| *B+* | 4 |
| *B* | 3 |
| *B−* | 2 |
| *CCC+* | 1 |
| *CCC* | 1 |
| *CCC−* | 1 |
| *CC* | 1 |
| *SD* | 1 |
| *D* | 1 |

**Table A.2:** Data Description

| Variable | Description | Source |
|---|---|---|
| Yields | 10-year sovereign Portuguese government bond yields | Bank of Portugal |
| Ret | PSI 20 stock index returns | Reuters |
| Rat - Pt | Sovereign credit rating: Portugal | Standard & Poor's |
| Rat-It | Sovereign credit rating: Italy | Standard & Poor's |
| Rat-Sp | Sovereign credit rating: Spain | Standard & Poor's |
| Rat-Gr | Sovereign credit rating: Greece | Standard & Poor's |
| Rat-Ir | Sovereign credit rating: Ireland | Standard & Poor's |

**Table A.3:** Data Descriptive Statistics

| Variable | Mean | Median | Max | Min | Std Dev | Skewness | Kurtosis | Observations |
|---|---|---|---|---|---|---|---|---|
| Yields | 0.050 | 0.045 | 0.140 | 0.031 | 0.019 | 2.699 | 10.398 | 3140 |
| Ret | 0.000 | 0.000 | 0.102 | $-0.104$ | 0.012 | $-0.157$ | 11.394 | 3140 |
| Rat-Pt | 13.719 | 14 | 15 | 6 | 1.895 | $-1.946$ | 6.043 | 3140 |
| Rat-It | 16.176 | 16 | 17 | 12 | 0.721 | $-0.039$ | 1.574 | 3140 |
| Rat-Sp | 13.912 | 14 | 15 | 10 | 0.942 | $-0.665$ | 3.539 | 3140 |
| Rat-Gr | 10.820 | 12 | 13 | 1 | 2.837 | $-2.406$ | 8.077 | 3140 |
| Rat-Ir | 15.951 | 17 | 17 | 10 | 1.942 | $-2.186$ | 6.718 | 3140 |

**Table A.4:** Rating Transformations

The $q_{ji}$ quantities represent the $i-th$ cut-off value for the $j-th$ country's rating ($V_{jt}$ ) defined in the model (5.26).

| $q_{ji}$ | Portugal | Italy | Spain | Greece | Ireland |
|---|---|---|---|---|---|
| $q_{j1}$ | 8 | 14 | 12 | 7 | 10 |
| $q_{j2}$ | 11 | 15 | 13 | 10 | 15 |
| $q_{j3}$ | 14 | 16 | 14 | 12 | 16 |

**Table A.5:** Unit Root Tests

| $H_o$ : | Deterministic Regressors | Test | $p-value$ |
|---|---|---|---|
| yields has a unit root | constant | Dickey - Fuller | 1.537 (0.959) |
| | constant | Phillips-Perron | 1.570 (0.955) |
| | linear trend | Dickey - Fuller | 0.586 (0.995) |
| | linear trend | Phillips-Perron | 0.612 (0.979) |

$t-statistics$ are reported. $p-values$ in parentheses.

The lag lenght criteria for the DF tests was the Schwartz info criteria

The spectral estimation method for PP tests was the Bartlett kernel and the bandwidth was the Newey-West one

**Table A.6:** ARCH Type Effects Test

We fit two preliminary autoregressive processes for level yields and for first differenciated yields and keep the respective residual series. In both cases we have evidence of conditional heteroskedasticity. We have used six lags in the tests

| | $yields-level$ | $yields-differenced$ |
|---|---|---|
| $F-Stat \sim F_{(6,3127)}$ | 5.6859 (0.000) | 6.1996 (0.0000) |
| $nR_u^2 \xrightarrow{d} \chi^2_{(6)}$ | 33.8226 (0.0000) | 36.8427 (0.0000) |

$F-Statistic$ and $\chi^2-Statistic$ are reported. $p-values$ in parentheses.

The degrees of freedom of the $F-Stat$ for the second case are $(6,3126)$

**Table A.7:** Granger Causality Tests

Table A.7 reports the Granger causality tests. The null hypothesis in this case is: each variable in the rows "does not Granger cause" each variable in the columns. $H_o$ is clearly rejected in all situations.

| $H_o$ : does not Granger Cause | Yields - levels | Yields - differences |
|---|---|---|
| $Rat - Pt$ | 23.4664 (0.000) | 23.2109 (0.000) |
| $Rat - It$ | 2.13043 (0.019) | 2.26119 (0.013) |
| $Rat - Sp$ | 10.5018 (0.000) | 10.7917 (0.000) |
| $Rat - Gr$ | 24.6176 (0.000) | 23.6902 (0.000) |
| $Rat - Ir$ | 5.03362 (0.000) | 3.73089 (0.000) |

$LR - Statistic$ is reported. $P - values$ in parentheses. We have used 6 lags.

**Table A.8:** Forecast Errors: Granger Causality

Table (A.8) suggests that the dummy variables vector $\boldsymbol{z}_t'$ defined in (5.28) Granger cause the Portuguese sovereign bond (differenciated) yields. Model 1 and Model 2 are designated models (5.29) and (5.28) respectively. It should be noted that the inclusion of the dummies in the model improves the forecast error of the dependent variable, given that the forecast errors of Model 2 are smaller than those of Model 1.

| $h$ | $Model$ | $MAPE$ | $MAE$ | $RMSE$ |
|---|---|---|---|---|
| 120 | 1 | 0.0724 | 0.0091 | 0.0119 |
| | 2 | 0.0546 | 0.0064 | 0.0072 |
| 90 | 1 | 0.0736 | 0.0094 | 0.0122 |
| | 2 | 0.0515 | 0.0066 | 0.0085 |
| 60 | 1 | 0.0807 | 0.0104 | 0.0128 |
| | 2 | 0.0751 | 0.0097 | 0.0121 |

$h$ represents the out-of-sample dimension

**Table A.9:** GARCH Estimation: Heteroskedasticity Test

Table (A.9) reports the heteroskedasticity test of the standardized residuals of the estimation of the model (5.28). The results of the estimations are displayed in table (5.5). We cannot reject the null hypothesis of conditional homoskedasticity.

| Statistic | Yields - differences |
|---|---|
| $F - Stat \sim F_{(6,3127)}$ | 0.9202 (0.4791) |
| $nR_u^2 \xrightarrow{d} \chi^2_{(6)}$ | 5.5237 (0.4786) |

$F - Statistic$ and $\chi^2 - Statistic$ are reported. $p - values$ between parentheses.