Statistics and Probability Letters 90 (2014) 108–113

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Combining a regression model with a multivariate Markov chain in a forecasting problem

Bruno Damásio, João Nicolau *

Universidade de Lisboa, ISEG, CEMAPRE, Portugal

ARTICLE INFO

Article history:
Received 21 November 2013
Received in revised form 17 February 2014
Accepted 27 March 2014
Available online 1 April 2014

Keywords:
Multivariate Markov chain
Higher-order Markov chain
Forecasting

ABSTRACT

This paper proposes a new concept: the usage of Multivariate Markov Chains (MMC) as co-variates. Our approach is based on the observation that we can treat possible categorical (or discrete) regressors, whose values are unknown in the forecast period, as an MMC in order to improve the forecast error of a certain dependent variable. Hence, we take advantage of the information about the past state interactions between the MMC categories to forecast the categorical (or discrete) regressors and improve the forecast of the actual dependent variable.

1. Introduction

Consider a simple regime-switching model

$$y_t = \beta x_t + \delta z_t + u_t$$

where z_t is a latent dummy variable that evolves over time according to a homogeneous Markov chain (i.e. $P(z_t = i_0 | z_{t-1} = i_1), i_0, i_1 = 0, 1$). This model and further refinements have been extensively studied in the literature (see Hamilton (1989)). In some circumstances the z_t variable may be observable, and in this case standard methods of estimation of β and δ apply. However, forecasting y_t may raise some difficulties because z_t (which is assumed to be a random variable) is not observable in the forecasting period (to simplify one assumes that x_t is a dynamic term, e.g. AR(1), or a simple trend). In this case a probabilistic structure is needed for z_t, for example a Markov chain, as in regime-switching models. In this paper we analyze the forecasting problem when the y_t variable depends on $s > 1$ discrete or categorical variables (observable during the estimation period), whose dependencies are governed by a multivariate Markov chain. This approach is new in the literature and the closest model to ours is perhaps the regime-switching one cited above. However, in contrast to regime-switching models which only deal with univariate Markov chains, usually with few states (in most cases with two or three states), given the complexity of the estimation procedures, our model is able to involve many "z_t" variables, with multiple states, thanks to the MTD-probit specification as we explain later on.

To be more precise, this paper considers the forecasting of a time series (y_t) that depends on quantitative variable(s) (x_t) and on s discrete or categorical variables, $(S_{1t}, \ldots, S_{st})$ where S_{jt} $(j = 1, \ldots, s)$ can take on values in the finite set $\{1, 2, \ldots, m\}$. We assume that S_{jt} depends on the previous values of $S_{1t-1}, \ldots, S_{jt-1}, \ldots, S_{st-1}$, and this dependence is well modeled by a first-order MMC. However, S_{jt} can also depend on some explanatory variables lagged over more than one period — our approach may in fact be viewed as a higher-order MMC (e.g. we may take S_{jt-1} as S_{t-j}, and in this case we would have an

* Correspondence to: School of Economics and Management (ISEG), Rua do Quelhas 6, 1200-781 Lisboa, Portugal. Tel.: +351 21 3925876; fax: +351 21 3922782.

E-mail address: nicolau@iseg.utl.pt (J. Nicolau).

http://dx.doi.org/10.1016/j.spl.2014.03.026
0167-7152/

s-order Markov chain). We propose using MMC as covariates in a regression model in order to improve the forecast error of a certain dependent variable, provided it is caused, in the Granger sense, by the MMC. Traditionally, and so far, the published literature only addresses the MMC as an end in itself. Here we take advantage of the information about the past state interactions between the MMC categories to forecast the dependent variable more accurately. As far as we know this forecasting problem has not yet been analyzed in the literature.

To form a regression model relating y_t to the categorical variables, we convert the S_{jt} categories into a set of dummy variables as follows:

$$z_{jkt} = \mathcal{I}_{\{S_{jt}=k\}} \tag{1.1}$$

where $\mathcal{I}_{\{.\}}$ is the indicator function, $\mathcal{I}_{\{S_{jt}=k\}} = 1$ if $S_{jt} = k$ and 0 otherwise. The proposed methodology also supports the event where S_{jt} is a discrete variable with state space $\{1, 2, \ldots, m\}$ (say), in which case no dummy variables are needed.

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

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\delta} + u_t \tag{1.2}$$

where:

- $\boldsymbol{x}_t'$ may be a vector of both deterministic and stochastic components, like AR(1) or other $\mathcal{F}_{t-1}$ or $\mathcal{F}_t$ measurable predetermined terms. Here $\mathcal{F}_t$ represents the information available at time t, i.e. the σ-algebra generated by all events up to time t.
- $\boldsymbol{z}_t'$ is a vector of dummy variables z_{kjt}, concerning the MMC, defined in (1.1).
- $\{u_t\}$ is a white noise process mean independent of $\boldsymbol{x}_t'$ and $\boldsymbol{z}_t'$. We do not assume any distribution for u_t.

To forecast y_{t+h} we use the best predictor according to the expected squared forecast error:

$$E\left(y_{t+h} \middle| \mathcal{F}_t\right) = E\left(\boldsymbol{x}_{t+h}' \middle| \mathcal{F}_t\right)\boldsymbol{\beta} + E\left(\boldsymbol{z}_{t+h}' \middle| \mathcal{F}_t\right)\boldsymbol{\delta} \tag{1.3}$$

given the exogeneity of the disturbance term, i.e. $E\left(u_t \middle| \mathcal{F}_{t-1}\right) = 0\ \forall t$.

To illustrate, suppose that we have two categorical variables ($s = 2$) and each categorical datum takes on values in the set $\{1, 2, 3\}$, i.e. $m = 3$. Unwinding the vector $\boldsymbol{z}_t'$ and the vector $\boldsymbol{\delta}$ it follows that

$$y_{t+h} = \boldsymbol{x}_{t+h}'\boldsymbol{\beta} + \delta_{11}\mathcal{I}_{\{S_{1t}=1\}} + \delta_{12}\mathcal{I}_{\{S_{1t}=2\}} + \delta_{21}\mathcal{I}_{\{S_{2t}=1\}} + \delta_{22}\mathcal{I}_{\{S_{2t}=2\}} + u_t \tag{1.4}$$

where S_{jt} represents the j-th categorical series of the MMC (notice that the dummy variable trap is avoided with this specification). Since the values of S_{jt+h} are unknown in the forecasting periods, i.e. for $h \geqslant 1$, we explore possible dependencies between S_{jt+h} and past values of S_{1t+h} and S_{2t+h} using an MMC approach, to predict S_{jt+h}, and consequently, y_{t+h}. If both $S_{1t}$and S_{2t} are discrete variables, the regression equation is simpler:

$$y_{t+h} = \boldsymbol{x}_{t+h}'\boldsymbol{\beta} + \delta_1 S_{1t+h} + \delta_2 S_{2t+h} + u_t. \tag{1.5}$$

From Eqs. (1.4) or (1.5), it is clear that to forecast y_{t+h} one needs to evaluate $P\left(S_{jt+h} = k \middle| \mathcal{F}_t\right)$, for $k = 1, 2, \ldots, s$. To keep these expressions simple, we make the following assumptions:

Assumption 1.1. First order MMC.

$$P\left(S_{jt} = k \middle| \mathcal{F}_{t-1}\right) = P\left(S_{jt} = k \middle| S_{1t-1} = i_1, \ldots, S_{st-1} = i_s\right). \tag{1.6}$$

That is, S_{jt} given $\{S_{1t-1}, \ldots, S_{st-1}\}$ is independent of any other variables in $\mathcal{F}_{t-1}$.

Assumption 1.2. Homogeneous MMC.

We have a homogeneous MMC in the sense that

$$P\left(S_{jt} = k \middle| S_{1t-1}, \ldots, S_{st-1}\right) = P\left(S_{jt+h} = k \middle| S_{1t+h-1}, \ldots, S_{st+h-1}\right). \tag{1.7}$$

Assumption 1.3. Contemporaneous needless terms.

S_{jt} is independent of $\left\{S_{1t}, \ldots, S_{j-1t}, S_{j+1t}, \ldots, S_{st}\right\}$ given $\{S_{1t-1}, \ldots, S_{st-1}\}$, i.e.

$$\begin{aligned} &P\left(S_{jt} = k \middle| S_{1t} = i_1, \ldots, S_{j-1t} = i_{j-1}, S_{j+1t} = i_{j+1}, \ldots, S_{st} = i_s, S_{1t-1}, \ldots, S_{st-1}\right) \\ &\quad = P\left(S_{jt} = k \middle| S_{1t-1}, \ldots, S_{st-1}\right). \end{aligned} \tag{1.8}$$

To obtain the forecast of y_{t+h} we need to calculate $E\left(\boldsymbol{x}_{t+h}' \middle| \mathcal{F}_t\right)$ and $E\left(\boldsymbol{z}_{t+h}' \middle| \mathcal{F}_t\right)$. It is assumed the former expression is known, hence we focus on the latter expression. A generic element of $E\left(\boldsymbol{z}_{t+h}' \middle| \mathcal{F}_t\right)$ is $E\left(\boldsymbol{z}_{kj,t+h} \middle| \mathcal{F}_t\right)$ which, by Assumption 1.1, can be written as

$$\begin{aligned} E\left(\boldsymbol{z}_{kj,t+h} \middle| \mathcal{F}_t\right) &= P\left(\boldsymbol{z}_{kj,t+h} = 1 \middle| \mathcal{F}_t\right) = P\left(S_{j,t+h} = k \middle| \mathcal{F}_t\right) \\ &= P\left(S_{j,t+h} = k \middle| S_{1t} = i_1, \ldots, S_{st} = i_s\right). \end{aligned} \tag{1.9}$$

We use the MMC theory to estimate the expression (1.9), which ultimately leads to the expressions $E\left(\boldsymbol{z}_{t+h}' \middle| \mathcal{F}_t\right)$ and $E\left(y_{t+h} \middle| \mathcal{F}_t\right)$. We briefly cover the main aspects of MMC estimation theory in the next section.