Journal of Multivariate Analysis 131 (2014) 209–228

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On binary and categorical time series models with feedback

Theodoros Moysiadis *, Konstantinos Fokianos

Department of Mathematics & Statistics, University of Cyprus, Cyprus

ARTICLE INFO

Article history:
Received 14 July 2013
Available online 21 July 2014

AMS subject classifications:
primary 62M10
secondary 62F12
62J12

Keywords:
Autocorrelation
Categorical data
Hidden Markov models
Latent process
Logistic regression
Multinomial regression
Nominal data
Prediction
Weak dependence

ABSTRACT

We study the problem of ergodicity, stationarity and maximum likelihood estimation for multinomial logistic models that include a latent process. Our work includes various models that have been proposed for the analysis of binary and, more general, categorical time series. We give verifiable ergodicity and stationarity conditions for the analysis of such time series data. In addition, we study maximum likelihood estimation and prove that, under mild conditions, the estimator is asymptotically normally distributed. These results are applied to real and simulated data.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Motivated by the work of Russell and Engle [36,37], who proposed a new approach to model financial transactions data, we study the problem of ergodicity, stationarity and maximum likelihood estimation for the so called autoregressive conditional multinomial (ACM) model with the logistic link, and its generalizations. More specifically, Russell and Engle [36,37] develop a model for the joint distribution of discrete price changes and durations conditional on the observed history. This joint distribution of both processes is decomposed into the product of the conditional density of the mark and the marginal density of the arrival times, both conditioned on the past filtration of the joint information set. Modeling of the conditional density of the mark is accomplished by the ACM model of order (l_1, l_2), which is a general class, contained in the class introduced by model (8), when $l_1 = l_2 = 1$. The authors examine theoretical properties of the model. However, the issue of obtaining sufficient conditions for the transaction price process to be stationary and ergodic, has not been investigated thoroughly in the literature, to the best of our knowledge. We address the problem of stationarity and ergodicity for the general case of model (8); see Theorem 1. The goal of this contribution is to provide the necessary conditions, under which the MLE estimator is asymptotically normally distributed; see Lemma 1 and Theorem 2. One version of the ACM model is based on multinomial regression; a method with generalizes naturally the standard logistic regression.

An alternative approach to the ACM class of models is based on the probit link function. Such autoregressive models have been considered by Zeger and Qaqish [41], Rydberg and Shephard [38], Kauppi and Saikkonen [27], among others. The work by de Jong and Woutersen [10] provides asymptotic results for the case of the dynamic probit model (4).

* Corresponding author.
E-mail addresses: moysiadis.theodoros@gmail.com, moysiadis.theodoros@ucy.ac.cy (T. Moysiadis), fokianos@ucy.ac.cy (K. Fokianos).

http://dx.doi.org/10.1016/j.jmva.2014.07.004
0047-259X/© 2014 Elsevier Inc. All rights reserved.

Our approach is based on the theory of generalized linear models, see [31], and in particular our focus is on the multinomial distribution. It is an elementary exercise to show that the multinomial distribution belongs to the multivariate exponential family and as such, the theory of generalized linear models can be applied for modeling various types of categorical data; nominal, interval and scale. In this contribution we will be working with nominal data and therefore the multinomial logistic model is the natural candidate for model fitting; see [24,14], [28, Ch. 3], [16] for further discussion on modeling issues regarding categorical data.

We note that Markov chains provide a simple but important example of categorical time series where lagged values of the response are important in determining its future states. Markov modeling in the context of categorical time series however, can be problematic for two reasons. First, as the order of the Markov chain increases so does the number of free parameters; in fact, the number of free parameters increases exponentially fast. In addition, the assumption of Markov property requires the specification of the joint dynamics of the response and any possible covariates observed jointly; such a specification might not be possible, in general.

We will be studying models for binary and, more generally, categorical time series, which are driven by a latent process or a feedback mechanism. This type of models is quite analogous to GARCH models – see [3] – but they are defined in terms of conditional log-odds instead of conditional variances. Such feedback models make possible low dimensional parametrization, yet they can accommodate quite complicated data structures. Paradigms of feedback models, in the context of count time series for example, have been studied recently by Fokianos et al. [17], Franke [20], Fokianos and Tjøstheim [18], Neumann [32], Fokianos and Tjøstheim [19] and Doukhan et al. [11]. In particular, we note that this contribution is closer to the modeling approach suggested by Fokianos and Tjøstheim [18], because the main idea is essentially to use the canonical link process to model the observed data. Several other models for the analysis of categorical data have been studied; see the books by Joe [22] and MacDonald and Zucchini [30] and the recent articles by Biswas and Song [2] and Weiß [40].

The outline of the paper is as follows. Section 2 puts forward the main model that we consider and discusses some of its properties. Section 3 develops results regarding its probabilistic properties. Section 4 discusses maximum likelihood estimation. In Section 5, we verify all the theoretical results via a simulation study for two special cases of model (8) that are of specific interest that is model (3) and model (7). Finally, in Section 6 we perform a real data analysis, where we give additional motivation for financial applications. An Appendix contains the proofs of all theoretical results.

2. Dynamic modeling of binary and categorical time series

We will be interested on a categorical time series, say $\{\tilde{Y}_t,\ t = 1, \dots, N\}$, where N denotes the sample size. Let m be the number of possible categories. This means that for each t, the possible values of $\tilde{Y}_t$ are $1, 2, \dots, m-1, m$, corresponding to the first, second category and so on. In general, and especially for nominal data, the aforementioned assignment of integer values to the categories is rather arbitrary. It is usually made as a matter of convenience, and it should be clear that such an assignment is not unique. However, it is useful to note, that regardless of any assignment, the tth observation of a categorical time series, can be expressed by the vector $Y_t = (Y_{1t}, Y_{2t}, \dots, Y_{qt})^T$ of length $q = m - 1$ with the following elements

$$Y_{jt} = \begin{cases} 1, & \text{if the } j\text{th category is observed at time } t, \\ 0, & \text{otherwise,} \end{cases}$$

for $t = 1, 2, \dots, N$ and $j = 1, 2, \dots, q$. Consider an increasing sequence of σ-fields, say $\{\mathcal{F}_t\}_{t\geq 1}$, which will be specified in detail later. Denote further by $p_t = (p_{1t}, p_{2t}, \dots, p_{qt})^T$ the vector of conditional success probabilities given $\mathcal{F}_{t-1}$, that is

$$p_{jt} = P(Y_{jt} = 1|\mathcal{F}_{t-1}) = E(Y_{jt}|\mathcal{F}_{t-1}), \quad t = 1, 2, \dots, N,\ j = 1, 2, \dots, q.$$

It is clear that the last category is recovered by the correspondence

$$Y_{mt} = 1 - \sum_{j=1}^{q} Y_{jt} \quad \text{and} \quad p_{mt} = 1 - \sum_{j=1}^{q} p_{jt}.$$

In what follows d, A, B are generally unknown parameters. In fact, d is a real vector of dimension q and A, B are $q \times q$ real matrices. Even though these symbols will be employed for defining distinct models, their meaning will be clear from the context.

Our focus is on developing and studying models for categorical time series, which include a feedback mechanism or an unobserved hidden process. For instance, one can consider the following linear model

$$p_t = d + Ap_{t-1} + BY_{t-1}, \quad t \in \mathbb{Z}, \tag{1}$$

which can be viewed as a simple generalized linear model with identity link for categorical data. Such model was suggested by Russell and Engle [36] and Qaqish [34]. However model (1) cannot be applied easily to data, since its structure implies complicated restrictions on the parameters d, A and B. This is so, because each element of the vector of probabilities p_t should lie between zero and one. In fact, model (1) imposes more complicated restrictions on d, A, B, when covariates are under consideration.

Download English Version:

https://daneshyari.com/en/article/1145532

Download Persian Version:

https://daneshyari.com/article/1145532

Daneshyari.com