Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2311 – 2330
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables

Roberta Paroli[a,*], Luigi Spezia[b]

[a] *Dipartimento di Scienze Statistiche, Università Cattolica S.C., via Necchi 9, 20123 Milano, Italy*
[b] *Dipartimento di Statistica, Università Ca' Foscari, Venezia, Italy*

Received 3 October 2006; received in revised form 20 September 2007; accepted 20 September 2007
Available online 16 October 2007

Abstract

The Bayesian analysis of a non-homogeneous Markov mixture of periodic autoregressions with state-dependent exogenous variables is proposed to investigate a non-linear and non-Normal time series. It is performed within a Markov chain Monte Carlo framework, along four consecutive steps: the specification of the identifiability constraint; the selection of the exogenous variables which influence the observed process and the time-varying transition probabilities of the hidden Markov chain; the choice of the cardinality of the hidden Markov chain state-space and the autoregressive order; the estimation of the parameters. The selection of the exogenous variables is performed in the complex case of correlation between variables, by means of a new procedure. An application for relating the hourly mean concentrations of sulphur dioxide with six meteorological variables, recorded for three years by an air pollution testing station located in the lagoon of Venice (Italy), is presented. The reconstruction of the sequence of the hidden states, the restoration of the missing values occurring within the observed series, the description of the periodic component are also given.

Keywords: Hidden markov chain; Marginal likelihood; Metropolized-Kuo–MallicK; Permutation sampling; Regime switching

1. Introduction

Markov mixture of autoregressions have been introduced in econometric literature by Hamilton under the name of time series subject to changes in regime (Hamilton, 1990) or Markov-switching time series models (Hamilton, 1996), when different autoregressions, each one depending on a latent regime, alternate according to the regime switching, which is driven by an unobserved homogeneous Markov chain (see also Hamilton, 1989, 1993). Krolzig (1997), Kim and Nelson (1999), Franses and van Dijk (2000) provide generalizations and applications for a wide range of non-linear and non-Normal economic and financial time series; see also the recent exhaustive monography by Frühwirth-Schnatter (2006). The name Markov mixtures models dates back to Chib (1996).

* Corresponding author. Tel.: +39 02 7234 2716; fax: +39 02 7234 2671.
E-mail address: roberta.paroli@unicatt.it (R. Paroli).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.020

The roots of these models can be found in three different class of models: the Markov switching regressions (Goldfeld and Quandt, 1973); the probabilistic functions of Markov chain (Baum et al., 1970; Baum, 1972); the applications of hidden Markov models in speech recognition (Levinson et al., 1983; Rabiner, 1989; Juang and Rabiner, 1991).

Bayesian analysis of Markov mixture of autoregressions has been developed, among others, by McCulloch and Tsay (1994); Chib (1996); Billio et al. (1999), Frühwirth-Schnatter (2001, 2004).

We generalize Markov mixture of autoregressions in three directions, by assuming that the unobserved Markov chain is non-homogeneous and by joining a periodic component and some exogenous variables to the autoregressive processes. We give rise to non-homogeneous Markov mixtures of periodic autoregressions with exogenous variables and show that this class of model can be an appropriate instrument not only in econometrics, but also in environmental statistics, where such models have been rarely exploited (e.g., the hydrologic applications by Zucchini and Guttorp, 1991 or Hughes et al., 1999).

Here, the hidden Markov chain is non-homogeneous, that is the transition probabilities are time-varying and modelled via logit functions of explanatory variables. Diebolt et al. (1994) were the first who studied these models with a two-state Markov chain, whose transition probabilities depend on deterministic exogenous variables through a logistic function. Then, Raymond and Rich (1997) studied the same kind of models, linking the transition probabilities to the covariates through binary probit models. Bayesian estimation of time-varying transition probabilities has been done by Filardo and Gordon (1998). They considered a hidden Markov chain with two states only: this statement allows them to run Gibbs sampling, instead of introducing a Metropolis step in the numerical procedure. Hughes et al. (1999) introduced in hydrology a highly complex hidden Markov model with a non-homogeneous Markov chain defined on a general number of states, whose transition probabilities depend on stochastic exogenous variables, and state-dependent distributions described by autologistic models for multivariate binary data. Further references can be found in Frühwirth-Schnatter (2006, Section 12.6.1).

Exogenous, deterministic variables are also joined to the autoregressive structure of our model, following the ideas developed by McCulloch and Tsay (1994). Here, we propose that the number of exogenous variables affecting the dynamics of the hidden states and the observed process are state-dependent, that is the number and the kind of variables can be different in any state. This new feature makes necessary a tool for selecting exogenous variables.

In recent years, several methods have been proposed both in Bayesian and in frequentist literature for selecting exogenous variables in regression models: in Bayesian context, the most popular are Stochastic Search Variable Selection (SSVS) by George and McCullogh (1993), the unconditional priors Gibbs sampler (KM) by Kuo and Mallick (1998) and Gibbs Variable Velection (GVS) by Dellaportas et al. (2000). Alternative Bayesian solutions are provided, among others, by Nott and Green (2004), Nott and Kohn (2005), Schneider and Corcoran (2004). In the frequentist context advanced methods using parallel computational algorithms have been proposed recently: see, e.g., Gatu et al. (2007), Hofmann et al. (2007) and Kapetanios (2007). We started by considering the first three aforementioned Bayesian methods which are the most appropriate in solving our problem. The differences among them depend on how an indicator vector, which makes the selection, is involved in the model: in SSVS the indicator vector is involved through the prior of the regression coefficients; in KM it is part of the regression equation, while GVS is a hybrid of SSVS and KM. Given that the complexity of our model is high and the exogenous variables we consider are strongly correlated, SSVS, KM and GVS have poor performances; hence we propose the "Metropolization" of the algorithm of Kuo and Mallick (1998), giving rise to Metropolized-Kuo–MallicK (MKMK) method. The correct selection of exogenous variables by MKMK is due to the acceptance in block of both the coefficients and the indicators of the exogenous variables, because blocking increases the precision of the estimators and improves the chain mixing.

A periodic component is also added. In economic applications, models with seasonal variables are widely used, by considering dummy variables; here, given our environmental application, the periodicity is modelled by using harmonic components, whose amplitudes depend on the latent variables. Note that the periodicities cause the persistence of the hidden regime for all the times of the cycle.

Finally, we assume that missing observations can occur within the observed time series; they will be handled as unknown parameters and estimated numerically.

The outline of the paper is the following. In Section 2 we introduce the model, by using the Bayesian approach. Section 3 presents the new method we propose to perform variable selection. In Section 4 we analyse a real problem applying our procedures to an environmental time series: identifiability constraint specification, variable selection, model choice and parameter estimation will be performed.