MARKOV SWITCHING AUTOREGRESSIVE MODELS FOR THE ANALYSIS OF HYDROLOGICAL TIME SERIES

Roberta Paroli [1] and Luigi Spezia [2]

[1] Dipartimento di Scienze Statistiche, Università Cattolica SC, Milano, (e-mail: roberta.paroli@unicatt.it)

[2] Biomathematics & Statistics Scotland, Aberdeen, (e-mail: luigi@bioss.ac.uk)

ABSTRACT: Markov switching autoregressive models (MSARMs) are proposed here in order to tackle the non-linearity, non-Normality, non-stationarity, and long memory of time series in hydrology. Bayesian inference, model choice, and stochastic variable selection are performed numerically by Markov chain Monte Carlo algorithms. Hence, it is possible to efficiently fit the data, reconstruct the sequence of hidden states, restore the missing values, classify the observations into a few regimes, and select the covariates. The efficiency of MSARMs is demonstrated by applications to isotope signatures, turbidity measurements, and river temperature. Our proposal is very general and flexible and can be applied to any kind of environmental time series.

KEYWORDS: marginal likelihood, non-linearity, non-Normality, non-stationarity, variable selection

1 Introduction and Data

Hydrological time series are realisations of complex stochastic systems. A few issues need to be taken into account by the modellers: non-Normality, non-linearity, non-stationarity, and long memory. These issues can be analysed by Markov switching autoregressive models (MSARMs): a class of models that is a popular tool within the econometrics community to model complex time series but has been considered quite rarely in other disciplines, including environmental sciences. Among the few applications in hydrology, Birkel et al. (2012) modelled isotope signatures; Spezia et al. (2021) turbidity measurements and Spezia et al. (2023) water temperature. In this work, we investigate the dynamic variability of water temperature by analysing an hourly water temperature time series automatically recorded in the Gairn catchment, in the North-East of Scotland, for more than five years, along with some covariates affecting both the latent process (i.e. the time-varying transition probabilities

of the hidden Markov chain) and the observed process. The water temperatures is recorded hourly from 16th August 2012 to 23rd November 2017; the length of the series is 46224 points (i.e. 1926 days; more than five years), with 328 missing values (0.71% of the total number of observations). The range of the series is between -0.02°C and 22.41°C. The contemporary series of the hourly river flows is also available. We also studied an intermediate series of water temperature from 13th June 2014 to 31st August 2016 (19440 observations; 810 days; more than two years) with 209 missing values (1.08%) along with three covariates (flow, air temperature, rainfall). Finally, a short series was considered: 1200 observations (50 days) with no missing values recorded from 18th August to 6th October 2012 along with seven covariates (flow, air temperature, rainfall, wind speed, wind direction, radiation, soil temperature). The length of series of the exogenous variables was limited by the need to not have missing values in these deterministic sequences. This because missing values within the covariates might bias the results of our analyses.

We propose MSARMs within the Bayesian framework: inference, model choice, and variable selection are performed numerically by Markov chain Monte Carlo (MCMC) algorithms.

2 Model and Inference

MSARMs are pairs of discrete-time stochastic processes, one observed and one latent, or hidden. The hidden process is a finite-state Markov chain, whereas the observed process, given the Markov chain, is conditionally autoregressive. The dynamics of the observed process is driven by the dynamics of the latent one, so that each observation depends on the contemporary state of the Markov chain. By this theoretical structure, MSARMs allow: *i*) modelling non-linear and non-Normal time series by assuming that different autoregressions, each one depending on a hidden state, alternate according to the Markovian regime switching; *ii*) modelling a long-memory process; *iii*) classifying the observations into a small number of homogeneous groups, labelled as the regimes of the Markov chain.

Seven covariates were also incorporated into the model through both the hidden Markov chain (the transition probabilities are time-varying and dependent on the dynamics of these exogenous variables) and the observed process (the state-dependent exogenous variables are added to the past observations). Thus, we have time-varying means and autocovariances, and hence, a non-stationary model. The covariates are: river flow, air temperature, rainfall, wind speed, wind direction, radiation, and soil temperature. The data set is also

characterised by periodicities: the hourly temperatures vary according to the dynamics of the year and of the 24 hours of the day. Hence, both an annual and a state-dependent daily harmonic component are added to the observed process.

In the Bayesian framework, inference, model choice, and variable selection are performed numerically by MCMC algorithms. The basic scheme for parameter estimation in the observed process is Gibbs sampling which also allows both restoration of the missing values occurring within the series of observations and reconstruction of the sequence of hidden states. Two random walk Metropolis moves are used to estimate the parameters of the hidden Markov chain. Adding extra-steps to the basic Metropolis-within-Gibbs scheme we can also compute the marginal likelihood of the various competing models through the MCMC sample. This procedure enables us to select the best model within a set of models varying for the number of hidden states and the order of the autoregressive processes. The exogenous, deterministic variables appearing in the observed process may be different in any state and they may be different from those affecting the transition probabilities. The transition matrix is affected by two sets of covariates (possibly different from each other and different from those in the observed process), one for the transitions from a lower to a higher state, and another for the transition from a higher to a lower state. The selection of the covariates appearing in each state-dependent autoregression and in the transition matrix is performed stochastically through the Metropolised-Kuo-MallicK (MKMK) method, proposed by Paroli and Spezia (2008). In the case of non-homogeneous hidden Markov models and MSARMs with covariates, the MKMK method improves the performance of the competing techniques, especially when the explanatory variables are strongly correlated, and/or when the complexity of the model is high.

3 Results

The flexibility of the MSARMs is demonstrated by the three applications we considered. For the whole series with a single covariate, the best model has three hidden states and autoregressions of the fifth order. Thus, the non-linear model (three hidden states) worked better than the corresponding linear model (no hidden states). Flow is relevant in the observed process for two states only, while it is not selected in the hidden process and the Markov chain is homogeneous. For the intermediate series with three covariates, the best model has three hidden states and autoregressions of the sixth order. Again, the non-linear model (three hidden states) works better than the corresponding linear

model (no hidden states). Flow is relevant in the observed process for one state only, while air temperature is always selected both in the observed and the hidden process. For the short series with seven covariates, we obtain that the best model is the linear autoregression of the sixth order, with no hidden Markov chain behind. Air temperature, solar radiation, and soil temperature are the relevant variables to explain the water temperature dynamics. Thus, discharge is a proxy for water temperature modelling, when no other more directly related variables are available. In those situations, the latent states will help to model the long-term dynamics, in the absence of true predictors with a physical meaning. As we saw in the first two applications, the hidden regimes can have an interpretation related to the seasonality. In fact, the Markov chain shows an annual dynamics which anticipates the annual dynamics of the water temperatures. It is not surprising that for the short series (50 days, i.e. no annual periodicity) the model is not multi-state. It would be interesting to see what happens when considering the seven covariates on longer series, that is if the same covariates are selected in a non-linear model (i.e., with a multi-state hidden Markov chain). Our study provides a novel application of the suitability of the MSARMs in hydrological time series analysis and environmental sciences in general. We hope our work can motivate other scientists to approach MSARMs and give their highly structured time series a valuable interpretation.

References

BIRKEL C., PAROLI, R. SPEZIA L. DUNN S.M. TETZLAFF D., & SOULSBY, C. 2012. A new approach to simulating stream isotope dynamics using Markov switching autoregressive models. *Advances in Water Resources*, **26**, 308–316.

PAROLI, R., & SPEZIA, L. 2008. Bayesian variable selection in Markov mixture models. *Communications in Statistics - Simulation and Computation*, **37**, 25–47.

SPEZIA, L., GIBBS S. GLENDELL M. HELLIWELL R. PAROLI R., & POHLE, I. 2023. Bayesian analysis of high frequency water temperature time series through Markov switching autoregressive models. *Under second revision for Environmental Modelling & Software*.

SPEZIA, L., VINTEN A. PAROLI R., & STUTTER, M. 2021. An evolutionary Monte Carlo method for the analysis of turbidity high-frequency time series through Markov switching autoregressive models. *Environmetrics*, **32**, e2695.