Mathematics and Society
University of Isfahan
http://math-sci.ui.ac.ir

Journal of Mathematics and Society
ISSN (print): 2345-6493, ISSN (on-line): 2345-6507
Vol. 8 No. 4 (2023), pp. 1-21.

http://www.ui.ac.ir

PREDICTION OF PM2.5 POLLUTION IN TEHRAN AIR BASED ON TEMPERATURE AND PRESSURE USING MARKOVIAN REGIME-SWITCHING NON-PARAMETRIC ADDITIVE TRANSITIVE REGRESSION MODEL

MORTEZA AMINI

Abstract. In this paper, we introduce the Markovian regime-switching regression model, which is a graphical model based on the hidden Markov model. This model can be viewed as a clustered regression model, in which a Markov process models the transition from one cluster to another. These clusters are indeed the hidden states of the process, in the hidden Markov model, which are assumed to be a Markov process of order one. Besides, other assumptions of the hidden Markov model are assumed in this model, while the emission distribution is assumed to be the conditional distribution of the response given the covariates and the states. As an application of this model, the problem of prediction of PM 2.5 pollution in Tehran's air based on temperature and pressure during 2015-2017 using the Markovian regime-switching non-parametric additive transitive model, is considered and studied. Furthermore, the package hhsmm in R software, is introduced as a powerful tool for modeling the stated model.

1. Introduction

State-switching models are models in which the distribution of a sequence of observations (usually during a time interval) is controlled by a sequence of hidden states, such that the conditional distribution of observations given each state is different from that given others. Hidden Markov and semi-Markov Models [27] are the most common instances of state-switching models, in which the hidden state is a

Keywords: Prediction of air pollution, Markov process, R software, Bawm-Welch algorithm.
Communicated by Majid Asadi.
Article Type: Promotional Paper.
Received: 15/07/2023, Accepted: 04/11/2023, Published Online: 12-12-2023.
Cite this article: M. Amini, Prediction of PM2.5 pollution in Tehran air based on temperature and pressure using Markovian regime-switching non-parametric additive transitive regression model, *Journal of Mathematics and Society*, **8** no. 4 (2023) 1–21.
http://dx.doi.org/10.22108/MSCI.2023.138405.1593 .

Markov or semi-Markov process. Some other models, including the regime-switching models or Kalman-Filter model, are in this category. Various applications of such models are introduced by the researchers including, speech recognition [12], cognitive learning [24], brain performance modeling [15], modeling environmental processes [4, 5, 6], sequential analysis, reliability theory [7], biological analysis [8, 9, 27], and many other applications.

2. Main Results

A hidden Markov model is constructed by the following items:

(1) Transition Probability Matrix $\mathbf{\Gamma} = (\gamma_{ij})$, where

$$\gamma_{ij} = \Pr(S_{t+1} = j | S_t = i), \quad i, j = 1, \ldots, J,$$

such that

$$\sum_{i=1}^{J} \gamma_{ij} = 1, \quad j = 1, \ldots, J.$$

(2) Initial State Probability $\boldsymbol{\delta} = (\delta_j)$, where

$$\delta_j = \Pr(S_1 = j), \quad j = 1, \ldots, J; \qquad \sum_{j=1}^{J} \delta_j = 1.$$

(3) Observation distributions $f_1(y), \ldots, f_J(y)$, where

$$f_j(y) = \Pr(Y_t = y | S_t = j); \quad j = 1, \ldots, J,$$

which are also called state-dependent distribution or emission distribution. When y_t is a continuous random variable, $f_j(y)$ is a probability density function, which is usually a normal distribution or mixture of normal distributions.

The regime-switching regression model is introduced by [14] as follows:

$$y_t = x_t^T \beta_{s_t} + \sigma_{s_t} \epsilon_t, \tag{2.1}$$

in which $\{y_t\}$ is the sequence of responses, $\{x_t\}$ is the sequence of covariates, $\{\epsilon_t\}$ are sequence of (usually) i.i.d. normally distributed errors with zero mean and a variance equal to 1, and β_{s_t} and σ_{s_t} are the regression coefficients and the standard deviation of errors at state s_t, respectively.

A generalization of the model (2.1) to the the additive regime-switching regression model is introduced by [20] as follows:

$$y_t = \mu_{s_t} + \sum_{j=1}^{p} f_{j,s_t}(x_{j,t}) + \sigma_{s_t} \epsilon_t, \tag{2.2}$$

Letting $x_t = (y_{t-\ell}, \ldots, y_{t-L}, z_{t-\ell}, \ldots, z_{t-L})$, for lags $L > \ell \geq 1$ in (2.2), the non-parametric additive transitive regime-switching regression model is obtained.

http://dx.doi.org/10.22108/MSCI.2023.138405.1593

All models in this paper and all necessary tools for modeling, initialization, fitting, and prediction of these models are included in the R package hhsmm, which can be downloaded from https://cran.r-project.org/package=hhsmm. The reader is also referred to [3] for more information and examples about hhsmm package.

3. Summary of Proofs/Conclusions

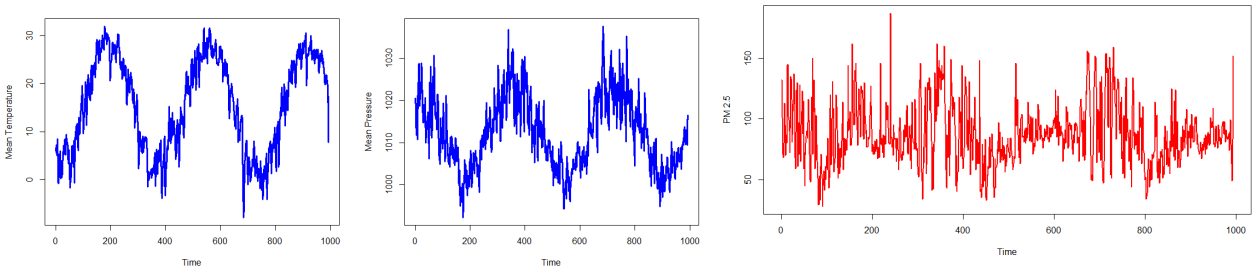

FIGURE 1. Temperature (left) pressure (middle) and PM2.5 (right) of Tehran city air during 2015-2017

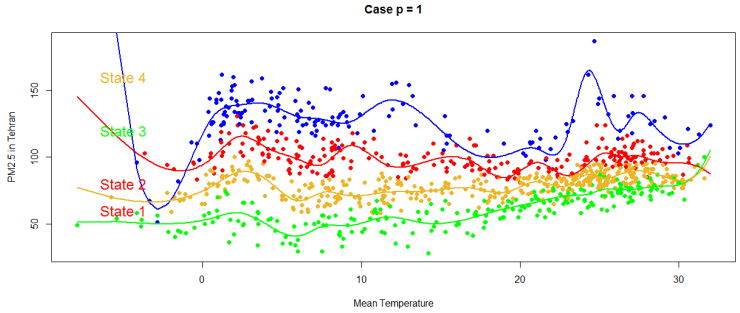

FIGURE 2. Prediction of PM2.5 in Tehran city air using regime-switching regression model, only based on the air temperature, in each of the four hidden states. The points in each state are colored by different colors and the curve of the prediction is drawn with the same color.

The data set for this paper is obtained from two sources. The AQI data set (PM2.5 values) are obtained from https://airnow.tehran.ir/, while the air temperature and pressure are obtained from Iran meteorological organization. Figure 1, shows the time-series plots of this data set.

To visualize the additive regime-switching regression model, we first consider only the temperature as the covariate in the model. Figure 2, presents the prediction of PM2.5 in Tehran city air using a nonparametric additive regime-switching regression model, only based on the air temperature, in each of the four hidden states. The points in each state are colored by different colors and the curve of the prediction is drawn with the same color. As one can see from this figure, the predictive curves are fairly fitted to the points in each state.

As a competitor model, we consider the single-state additive regression model. Figure 3, presents the result of the comparison of prediction precision of PM2.5 in Tehran city air, using two fitted models: the regime-switching regression model with four hidden states and the single state non-parametric additive regression model. The mean squared error in each model is presented in each plot. One can see that the non-parametric additive regime-switching regression model with four hidden states performs better than the single-state non-parametric additive regression model.

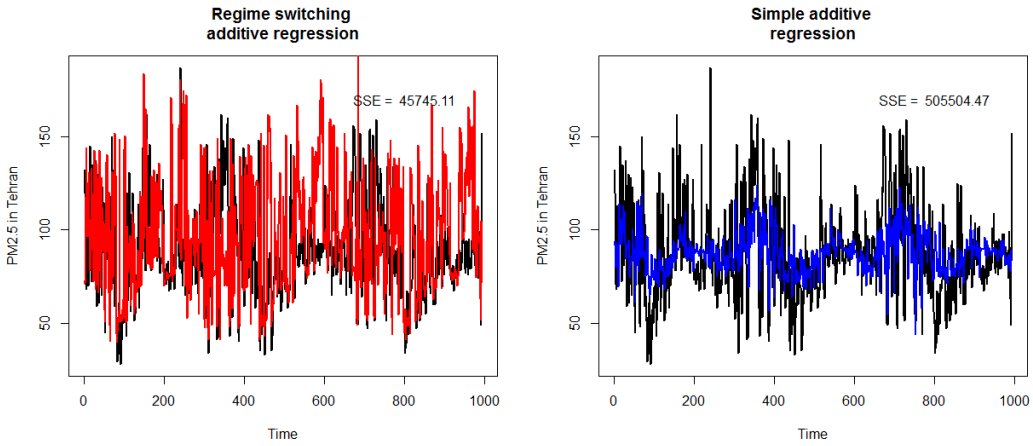

FIGURE 3. Comparison of prediction precision of PM2.5 in Tehran city air, using two fitted models: regime-switching regression model with four hidden states and the single state non-parametric additive regression model. The mean squared error in each model is presented in each plot.

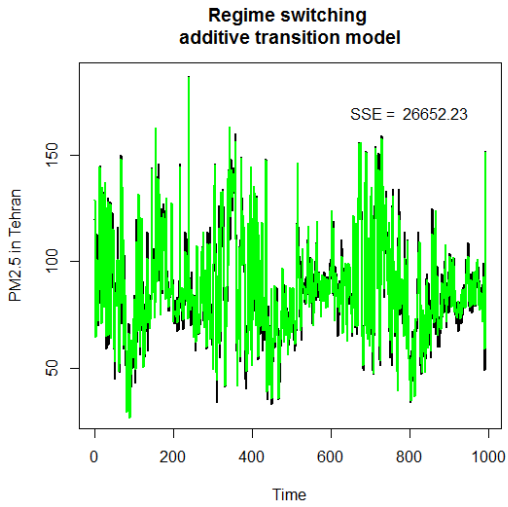

FIGURE 4. Prediction of PM2.5 in Tehran city air using transitive regime switching non-parametric additive regression model with 1-day lag. The mean squared error of the model is presented in the plot.

http://dx.doi.org/10.22108/MSCI.2023.138405.1593

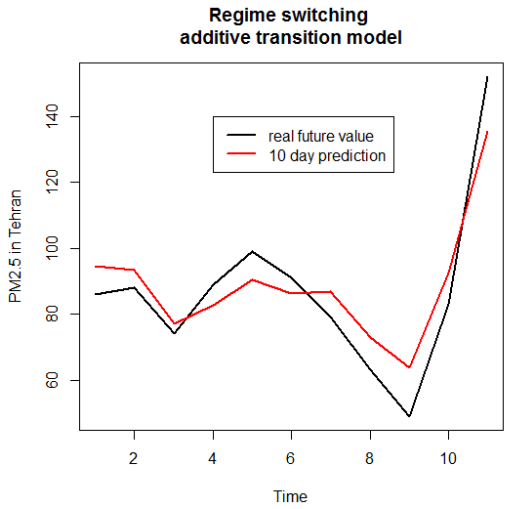

FIGURE 5. Out of sample prediction of PM2.5 in Tehran city air using transitive regime switching non-parametric additive regression model with 10 days lag. The mean squared error of the model is equal to 99.3.

Another introduced model is the additive transitive regime-switching regression model. Figure 4, shows the result of the prediction of PM2.5 in Tehran city air using a transitive regime switching non-parametric additive regression model with a 1-day lag. The mean squared error of the model is presented in the plot. One can see that this model performs better than the two other competitors.

Finally, the additive transitive regime-switching model is used for the prediction of the future values of PM2.5. Figure 5, presents the out-of-sample prediction of PM2.5 in Tehran city air using a transitive regime switching non-parametric additive regression model with 10 days lag. The mean squared error of the model is equal to 99.3. The result of the prediction is satisfactory.

REFERENCES

[1] T. Adam, R. Langrock and C. H. Weiß, Penalized estimation of flexible hidden Markov models for time series of counts, *Metron*, **77** (2) (2019) 87–104.

[2] R. M. Altman, Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting, *J. Amer. Statist. Assoc.*, **102** (477) (2007) 201–210.

[3] M. Amini, A. Bayat and R. Salehian, hhsmm: an *R* package for hidden hybrid Markov/semi-Markov models, *Comput. Statist.*, **38** (2022) 1283–1335.

[4] J. Bulla, F. Lagona, A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series, *J. Agric. Biol. Environ. Stat.*, **17** (4) (2012) 544–567.

[5] M. S. Bebbington, Identifying volcanic regimes using Hidden Markov Models, *Geophys. J. Int.*, **171** (2) (2007) 921–942.

[6] D. L. Borchers, W. Zucchini, M. P. Heide-Jorgensen, A. Cañadas and R. Langrock, Using hidden Markov models to deal with availability bias on line transect surveys, *Biometrics*, **69** (3) 703–713.

[7] F. Cartella, J. Lemeire, L. Dimiccoli and H. Sahli, Hidden semi-Markov models for predictive maintenance, *Math. Probl. Eng.*, (2015) 23 pp.

[8] G. A. Churchill, Stochastic models for heterogeneous DNA sequences, *Bull. Math. Biol.*, **51** (1) 79–94.

[9] R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press, 1998.

[10] Paul H. C. Eilers and B. D. Marx, Flexible smoothing with B-splines and penalties, *Statist. Sci.*, **11** (2) (1996) 89–121.

[11] Y. Guédon, Estimating hidden semi-Markov chains from discrete sequences, *J. Comput. Graph. Statist.*, **12** (3) (2003) 604–639.

[12] B. H. Juang and L. R. Rabiner, Hidden Markov models for speech recognition, *Technometrics*, **33** (3) (1991) 251–272.

[13] G. Kauermann, A note on smoothing parameter selection for penalized spline smoothing, *J. Statist. Plann. Inference*, **127** (1-2) (2005) 53–69.

[14] C. J. Kim, J. Piger and R. Startz, Estimation of Markov regime-switching regression models with endogenous switching, *J. Econometrics*, **143** (2) (2008) 263–273.

[15] R. Langrock, B. J. Swihart, B. S. Caffo, N. M. Punjabi and C. M. Crainiceanu, Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms, *Stat. Med.*, **32** (19) (2013) 3342–3356.

[16] R. Langrock, T. Michelot, A.Sohn and T. Kneib, Semiparametric stochastic volatility modelling using penalized splines, *Comput. Statist.*, **30** (2) (2015) 517–537.

[17] R. Langrock, T. Kneib, A. Sohn and S. L. DeRuiter, Nonparametric inference in hidden Markov models using P-splines, *Biometrics*, **71** (2) (2015) 520–528.

[18] R. Langrock, T. Kneib, R. Glennie and T. Michelot, Markov-switching generalized additive models, *Stat. Comput.*, **27** (1) (2017) 259–270.

[19] V. Leos-Barajas, E. J. Gangloff, T. Adam, R. Langrock, F. M. Van Beest, J. Nabe-Nielsen and J. M. Morales, Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures, *J. Agric. Biol. Environ. Stat.*, **22** (2017) 232–248.

[20] R. Langrock, T. Adam, V. Leos-Barajas, S. Mews, D. L. Miller and Y. P. Papastamatiou, Spline-based nonparametric inference in general state-switching models, *Stat. Neerl.*, **72** (3) (2018) 179–200.

[21] A. Maruotti, Mixed hidden markov models for longitudinal data: An overview, *Int. Stat. Rev.*, **79** (3) (2011) 427–454.

[22] S. Schliehe-Diecks, P. M. Kappeler, and R. Langrock, On the application of mixed hidden Markov models to multiple behavioural time series, *Interface focus*, **2** (2) (2012) 180–189.

[23] C. Sherlock, T. Xifara, S. Telfer and M. Begon, A coupled hidden Markov model for disease interactions, *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **62** (4) (2013) 609–627.

[24] I. Visser, M. E. J. Raijmakers and P. C. M. Molenaar, Fitting hidden Markov models to psychological data, *Sci. Program.*, **10** (3) (2002) 185–199.

[25] L. R. Welch, Hidden Markov models and the Baum-Welch algorithm, *IEEE Inf. Theory Soc. Newsl.*, **53** (4) (2003) 10–13.

[26] W. Zucchini, D. Raubenheimer and I. L. MacDonald, Modeling time series of animal behavior by means of a latent-state model with feedback, *Biometrics*, **64** (3) (2008) 807–815.

[27] W. Zucchini, I. L. MacDonald and R. Langrock, *Hidden Markov models for time series: an introduction using R*, Second edition, Monographs on Statistics and Applied Probability, **150**, CRC Press, Boca Raton, FL, 2016.

http://dx.doi.org/10.22108/MSCI.2023.138405.1593

Morteza Amini

Department of Statistics, School of Mathematics, Statistics, and Computer Science, Colledge of Science, University Tehran, Tehran, Iran

Email: morteza.amini ut.ac.ir