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.
© 2023 The Author(s)
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 $\beta_{s_t}$ and $\sigma_{s_t}$ 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

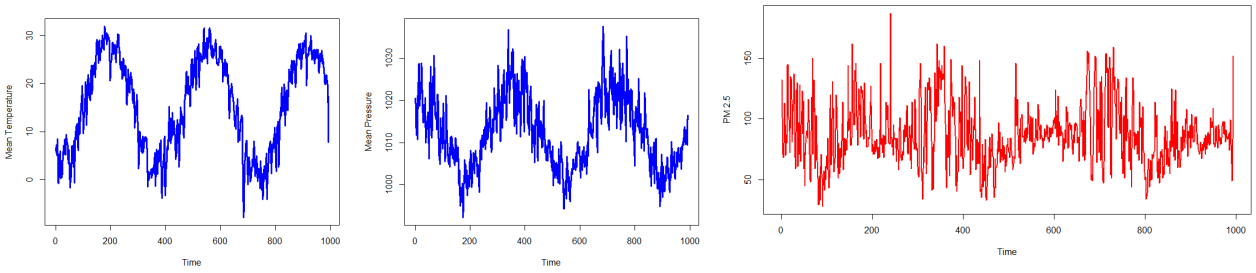



FIGURE 1. Temperature (left) pressure (middle) and PM2.5 (right) of Tehran city air during 2015-2017

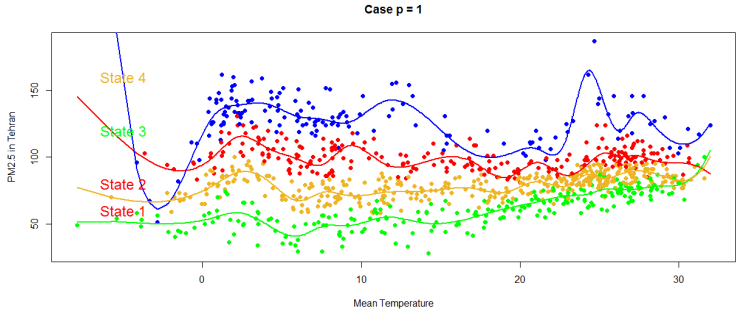



FIGURE 2. Prediction of PM2.5 in Tehran city air using regime-switching regression model, only based on the air temperature, in each of the four hidden states. The points in each state are colored by different colors and the curve of the prediction is drawn with the same color.

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.

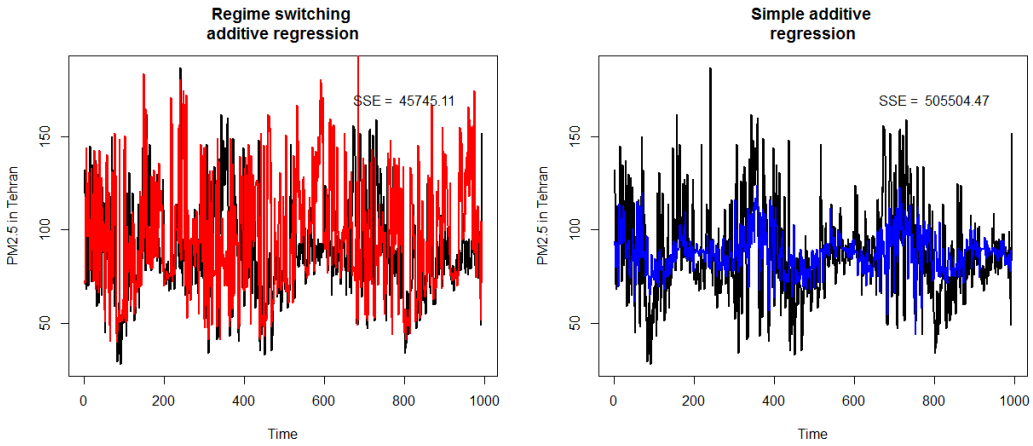



FIGURE 3. Comparison of prediction precision of PM2.5 in Tehran city air, using two fitted models: regime-switching regression model with four hidden states and the single state non-parametric additive regression model. The mean squared error in each model is presented in each plot.

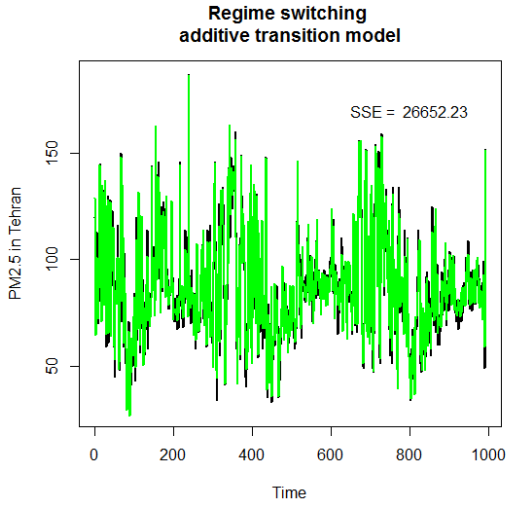



FIGURE 4. Prediction of PM2.5 in Tehran city air using transitive regime switching non-parametric additive regression model with 1-day lag. The mean squared error of the model is presented in the plot.


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

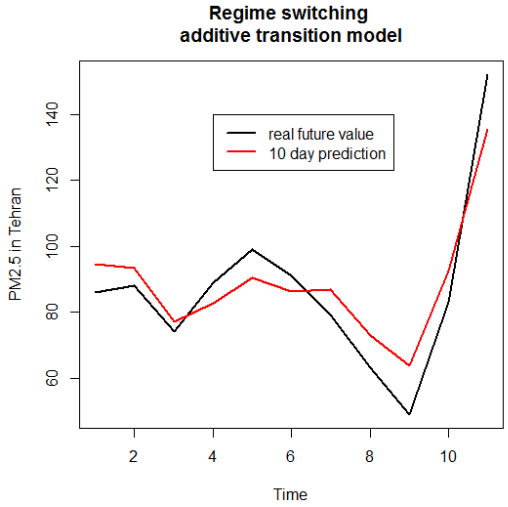



FIGURE 5. Out of sample prediction of PM2.5 in Tehran city air using transitive regime switching non-parametric additive regression model with 10 days lag. The mean squared error of the model is equal to 99.3.

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.

**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

ریاضی و جامعه
شاپا (چاپی): ۲۳۴۵-۶۴۹۳، شاپا(الکترونیکی): ۲۳۴۵-۶۵۰۷
جلد ۸، شماره ۴ (۱۴۰۲)، صص. ۱-۲۱
The Author(s) ۱۴۰۲ ©
http://www.ui.ac.ir
http://math-sci.ui.ac.ir

# پیش‌بینی ذرات معلق PM۲.۵ هوای تهران بر اساس دما و فشار هوا، با استفاده از مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی انتقالی

مرتضی امینی

**چکیده.** در این مقاله به معرفی مدل رگرسیون تغییر رژیم مارکوفی، که یک مدل گرافی بر مبنای مدل مارکوف پنهان است، می‌پردازیم. این مدل را می‌توان از زاویه دید دیگر گونه‌ای از مدل‌های رگرسیونی خوشه‌بندی شده در نظر گرفت؛ که در آن یک فرایند مارکوف پنهان، انتقال از خوشه‌ای به خوشه دیگر را مدل‌بندی می‌کند. این خوشه‌ها، در واقع همان وضعیت‌های پنهان در یک مدل مارکوف پنهان هستند که فرض می‌شود یک فرایند مارکوف از مرتبه اول باشند. همچنین فرض‌های پایه‌ای یک مدل مارکوف پنهان، در این مدل صادق هستند، با این تفاوت که توزیع مشاهده‌ها یک توزیع شرطی از متغیر پاسخ به شرط متغیرهای پیشگو و وضعیت پنهان است. به عنوان یک کاربرد این مدل، مسأله پیش‌بینی ذرات معلق PM۲.۵ در هوای شهر تهران بر اساس دما و فشار هوا و در فاصله سال‌های ۱۳۹۴ تا ۱۳۹۶، با استفاده از مدل رگرسیون تبدیل مارکوفی ناپارامتری جمعی انتقالی مطرح و بررسی شده است. همچنین بسته نرم‌افزاری hhsmm در نرم‌افزار R، به‌عنوان ابزاری قدرتمند برای مدل‌بندی مدل بیان شده، معرفی شده است.

## ۱. مقدمه

مدل‌های تبدیل وضعیت[1]، مدل‌هایی هستند که در آن‌ها توزیع دنباله‌ای از مشاهده‌ها (معمولا در طول زمان) توسط یک فرایند از وضعیت‌های پنهان کنترل می‌شود، به‌طوری‌که توزیع مشاهده‌ها در هر وضعیت تغییر می‌کند. مدل‌های مارکوف و نیمه مارکوف پنهان،[2][۲۷] متداول‌ترین نمونه از مدل‌های تبدیل وضعیتی هستند که در آن‌ها فرایند وضعیت پنهان یک فرایند مارکوف و یا نیمه مارکوف است. مدل‌های بسیار دیگری نیز مانند مدل‌های رگرسیونی تغییر رژیم مارکوفی و مدل کالمن فیلتر، نیز در این دسته جای می‌گیرند. کاربردهای گوناگونی از این مدل‌ها توسط پژوهشگران ارائه شده است که از آن جمله می‌توان به تشخیص گفتار [۱۲]، مدل‌سازی فعالیت مغز [۱۵]، یادگیری شناختی [۲۴]، مدل‌سازی فرایندهای زیست‌محیطی [۴، ۵، ۶]، تحلیل دنباله‌های زیستی [۸، ۹، ۲۷]، و قابلیت اعتماد صنعتی [۷]، اشاره کرد.

پژوهش‌های گوناگونی در زمینه مدل‌های تبدیل وضعیت برای توسعه ساختار این مدل‌ها انجام شده است، که از آن جمله می‌توان به مدل‌های نیمه مارکوف و ترکیبی پنهان [۱۱]، مدل‌های مارکوف پنهان مفصلی [۲۳]، مدل‌هایی با بازخورد از مشاهدات به فرایند وضعیت [۲۶]، یا فرایندهای وضعیت با ساختار سلسله مراتبی [۱۹]، اشاره کرد. همچنین برخی پژوهش‌ها نیز به

عبارات و کلمات کلیدی: پیش‌بینی آلودگی هوا، فرایند مارکوف، نرم افزار $R$، الگوریتم بام ولش.
دبیرتخصصی رابط: محمدرضا پوریای‌ولی
نوع مقاله: ترویجی
تاریخ دریافت: ۱۴۰۲/۰۴/۲۴ تاریخ پذیرش: ۱۴۰۲/۰۸/۱۳ تاریخ انتشار آنلاین: ۱۴۰۲/۰۹/۲۱
ارجاع به مقاله: م. امینی، پیش‌بینی ذرات معلق PM۲.۵ هوای تهران بر اساس دما و فشار هوا، با استفاده از مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی انتقالی، نشریه ریاضی و جامعه، ۸ شماره ۴ (۱۴۰۲) ۱-۲۱.
. http://dx.doi.org/10.22108/msci.2023.138405.1593

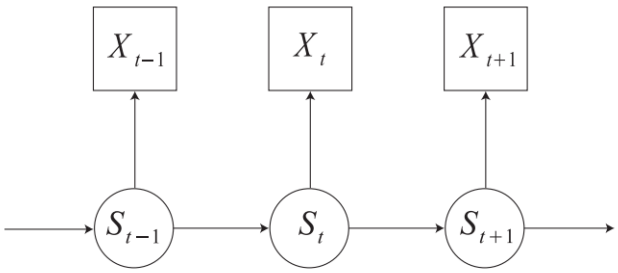
[1]state-switching models [2]Hidden Markov and semi-Markov models

استنباط در مدل‌های مارکوف پنهان آمیخته برای داده‌های طولی پرداخته‌اند [۲، ۲۱، ۲۲]، که در آن‌ها از اثرات تصادفی برای محاسبه ناهمگنی بالقوه در طول سری‌های زمانی چندگانه استفاده شده است.

مدل‌های تغییر رژیم مارکوفی یک مدل گرافی بر مبنای مدل مارکوف پنهان است، که آن‌ها را می‌توان از زاویه دید دیگر گونه‌ای از مدل‌های رگرسیونی خوشه‌بندی شده در نظر گرفت، که در آن یک فرایند مارکوف پنهان، انتقال از خوشه‌ای به خوشه دیگر را مدل‌بندی می‌کند. این خوشه‌ها در واقع همان وضعیت‌های پنهان در یک مدل مارکوف پنهان هستند که فرض می‌شود یک فرایند مارکوف از مرتبه اول باشند. همچنین فرض‌های پایه‌ای یک مدل مارکوف پنهان در این مدل صادق هستند، با این تفاوت که توزیع مشاهده‌ها یک توزیع شرطی از متغیر پاسخ به شرط متغیرهای پیشگو و وضعیت پنهان است. مدل رگرسیون در نظر گرفته شده در مدل‌های تغییر رژیم مارکوفی معمولاً یک مدل رگرسیون خطی پارامتری است. به‌تازگی، استفاده از روش‌های هموارسازی اسپلاین برای استنباط ناپارامتری در مدل‌های تبدیل وضعیت پیشنهاد شده است [۱۶، ۱۷، ۱۸، ۱]، با استفاده از این ابزار می‌توان مدل‌های رگرسیون ناپارامتری جمعی را به مدل‌های رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی تعمیم داد. در این مقاله، ما به معرفی مدل رگرسیون تغییر رژیم مارکوفی می‌پردازیم و از آن برای پیش‌بینی ذرات معلق PM۲.۵ در هوای شهر تهران، بر اساس دما و فشار هوا و در فاصله سال‌های ۱۳۹۴ تا ۱۳۹۶، بهره گرفته‌ایم.

ساختار این مقاله به‌شرح زیر است. در بخش دوم به معرفی مدل‌های مارکوف پنهان و رگرسیون تغییر رژیم مارکوفی می‌پردازیم. در بخش سوم به شیوه برآورد پارامترها در این مدل‌ها پرداخته می‌شود. در بخش چهارم مسأله رمزگشایی وضعیت‌های پنهان به کمک الگوریتم ویتربی، بیان می‌شود. در بخش ششم، بسته نرم‌افزاری hhsmm در نرم افزار R، به عنوان ابزاری کارآمد برای برازش مدل‌های بیان شده، معرفی می‌گردد. در نهایت در بخش هفتم، به تحلیل داده‌های آب و هوای تهران و بیان نتایج و همچنین مقایسه برخی مدل‌های رقیب خواهیم پرداخت.

## ۲. مدل‌های مارکوف پنهان و رگرسیون تغییر رژیم مارکوفی




**شکل ۱.** نمایش گرافی مدل مارکوف پنهان.

**Figure** 1: Graphical representation of hidden Markov model.

همان‌طور که گفتیم مدل مارکوف پنهان، شامل دو فرایند تصادفی است، که یکی از آن‌ها قابل مشاهده و دیگری پنهان است. فرایند مشاهده شده یک دنباله چون $\{Y_t\}_{t=1}^T$، و دنباله وضعیت‌های پنهان، یک دنباله مانند $\{S_t\}_{t=1}^T$ است. یک زنجیره مارکوف دارای $J$ وضعیت فرض می‌شود، که وضعیت در زمان $t$ تنها به‌وضعیت در زمان $t-1$ وابسته است و فرض می‌شود که هریک از مشاهده‌ها، از یکی از توزیع‌های وابسته به $J$ وضعیت تولید شده است. علاوه بر آن، هر مشاهده به شرط وضعیت پنهان زمان خود از سایر مشاهده‌ها مستقل فرض می‌شود، یعنی $Y_t$ به شرط $S_t$ از $Y_{t'}, t' \neq t$ مستقل است. یک مدل مارکوف پنهان ایستا با سه دسته پارامتر زیر مشخص می‌شود:


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

(۱) ماتریس احتمال انتقال[۳] $\mathbf{\Gamma} = (\gamma_{ij})$ ، که در آن

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

به‌طوری‌که

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

(۲) بردار احتمال‌های اولیه[۴] $\boldsymbol{\delta} = (\delta_j)$ ، که در آن

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

(۳) توزیع‌های مشاهده‌های $f_1(y), \ldots, f_J(y)$ که

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

که توزیع‌های وابسته به وضعیت[۵] یا توزیع‌های گسیل[۶] نیز نامیده می‌شوند. بدیهی است در صورتی‌که $y_t$ متغیر تصادفی مطلقاً پیوسته باشد، $f_j(y)$ یك تابع چگالی احتمال است. در این‌حالت $f_j(y)$ معمولاً توزیع نرمال[۷] یا توزیع نرمال آمیخته[۸] فرض می‌شود.

برای به‌دست آوردن تابع درستنمایی مدل مارکوف پنهان بایستی توزیع توأم مشاهده‌ها و وضعیت‌های پنهان را روی تمام مقادیر $S_1, \ldots, S_T$ جمع بست، که منجر به تابعی پیچیده شده و به‌دست آوردن برآوردهای بیشینه درستنمایی را دچار چالش می‌کند. برای رفع این مشکل و با توجه به پنهان بودن متغیرهای $S_1, \ldots, S_T$، برای برآورد پارامترها از الگوریتم (EM) استفاده می‌کنیم. این الگوریتم دارای دو گام (E)[۹] و گام (M)[۱۰] است. از آن‌جا که محاسبه امیدریاضی شرطی $S_t$ به شرط $y_1, \ldots, y_T$ به سادگی قابل انجام نیست. در گام E برای محاسبه امید ریاضی از الگوریتم پیشرو–پسرو[۱۱] استفاده می‌کنیم که ترکیب آن با الگوریتم EM به الگوریتم بام–ولش[۱۲] [۲۵] مشهور است. در این الگوریتم از احتمال‌های پیشرو و پسرو، برای محاسبه تابع درستنمایی و برآورد پارامترهای استفاده می‌شود.

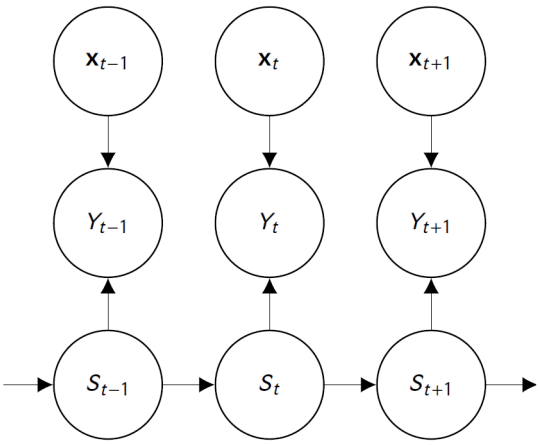
### ۱.۲. مدل‌های رگرسیون تغییر رژیم مارکوفی خطی و ناپارامتری جمعی.

مدل رگرسیون تغییر رژیم مارکوفی گاوسی در [۱۴] به‌صورت زیر معرفی شد:

$$y_t = x_t^T \beta_{s_t} + \sigma_{s_t} \epsilon_t, \tag{۱}$$

که در آن $y_t$ متغیر پاسخ، $x_t$ بردار متغیرهای پیشگو که می‌تواند شامل تأخیرهای زمانی $y_t$ نیز باشد (مدل مارکوف پنهان اتورگرسیو و رگرسیون انتقالی تغییر رژیم مارکوفی، بخش بعد را ببینید)، $s_t$ وضعیت پنهان، و $\epsilon_t$ خطاهای مدل رگرسیون است که فرض می‌شود مستقل و هم توزیع از توزیع نرمال با میانگین صفر و واریانس همگن برای $t = 1, \ldots, T$ باشند. تفاوت بین مدل تغییر رژیم مارکوفی (۱) و مدل مارکوف پنهان این است که به جای در نظر گرفتن تابع چگالی احتمال $y_t$ به شرط $s_t$ به عنوان توزیع گسیل، از توزیع شرطی $y_t$ به شرط $x_t$ و $s_t$ استفاده می‌شود. نمایش گرافی مدل رگرسیون تغییر رژیم مارکوفی در شکل ۲ نشان داده شده است. پارامترهای این مدل نیز مانند مدل مارکوف پنهان توسط الگوریتم EM و ترکیب آن با روش پیش‌رو–پس‌رو یعنی الگوریتم باوم–ولش برآورد می‌شوند.

[3] transition probability matrix [4] initial state probability [5] state-dependent distribution [6] emission distribution [7] normal distribution [8] mixture normal distribution [9] expectation [10] maximization [11] forward-backward algorithm [12] Baum-Welch algorithm




**شکل ۲.** نمایش گرافی مدل رگرسیون تغییر رژیم مارکوفی.

**Figure** 2: Graphical representation of regime switching regression model.

تعمیمی از مدل (۱) به مدل تغییر رژیم مارکوفی ناپارامتری جمعی زیر در [۲۰] چنین معرفی شده است:

$$y_t = \mu_{s_t} + \sum_{j=1}^{p} f_{j,s_t}(x_{j,t}) + \sigma_{s_t}\epsilon_t, \tag{۲}$$

که در آن $p$ $f_{j,s_t}(\cdot),\ j = 1,\ldots,p$ توابع رگرسیون نامعلوم برای $p$ متغیر پیشگو هستند. آن‌ها از B – اسپلاین‌های جریمه شده در تابع درستنمایی گاوسی برای برازش مدل (۲) در هر گام الگوریتم باوم–ولش استفاده کردند.

**۲.۲. مدل رگرسیون انتقالی تغییر رژیم مارکوفی.** یک حالت خاص از مدل‌های تغییر رژیم مارکوفی (۱) و (۲) مدل رگرسیون انتقالی تغییر رژیم مارکوفی است، که در آن قرار می‌دهیم $x_t = (y_{t-\ell},\ldots,y_{t-L},z_{t-\ell},\ldots,z_{t-L})$، که در آن تأخیرهای $1 \le \ell < L$ مشخص هستند و $z_t$ دنباله‌ای از متغیرهای پیشگو است. از این مدل می‌توان برای پیش‌بینی آینده متغیر پاسخ، بر اساس مقدارهای زمان حال خود متغیر پاسخ و برخی متغیرهای پیشگو در یک مدل رگرسیون تغییر رژیم مارکوفی، استفاده کرد.

## ۳. برآورد پارامترها توسط الگوریتم بام–ولش

همان‌طور که گفتیم برای برآورد پارامترهای مدل‌های مارکوف پنهان و رگرسیون تغییر رژیم مارکوفی، از ترکیب الگوریتم EM و الگوریتم پیش‌رو–پس‌رو استفاده می‌شود که به الگوریتم بام–ولش معروف است. در فاز E الگوریتم EM الگوریتم پیش‌رو–پس‌رو انجام می‌شود. گام پیشرو الگوریتم پیشرو–پسرو برای محاسبه $\Pr(y_1,\ldots,y_t,S_t=j)$ برای $t = 1,\ldots,T$ است. بنابراین در این گام احتمال‌های زیر محاسبه می‌شود:

$$\boldsymbol{\alpha}_t = (\alpha_t(1),\ldots,\alpha_t(J)); \quad \alpha_t(j) = \Pr(y_1,\ldots,y_t,S_t=j), \quad j = 1,\ldots,J$$

داریم

$$\Pr(y_1,\ldots,y_t) = \sum_{j=1}^{J} \alpha_t(j).$$


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

از $\alpha_t(j)$ می‌توان برای محاسبه احتمال‌های شرطی زیر استفاده کرد:

$$\Pr(S_t = j|y_1, \ldots, y_t) = \frac{\alpha_t(j)}{\sum_{i=1}^{J} \alpha_t(i)}.$$

می‌توان نشان داد که فرمول بازگشتی پیشرو زیر برای محاسبه بردار $\boldsymbol{\alpha}_t$ برقرار است:

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(y_t); \quad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(y_t), \quad t = 2, \ldots, T, \tag{۳}$$

که در آن $\mathbf{P}(y_t) = \text{diag}(f_1(y_t), \ldots, f_J(y_t))$ است.

گام پسرو الگوریتم پیشرو–پسرو و برای محاسبه احتمال‌های زیر است

$$\boldsymbol{\beta}_t = (\beta_t(1), \ldots, \beta_t(J)); \quad \beta_t(j) = \Pr(y_{t+1}, \ldots, y_T|S_t = j), \quad j = 1, \ldots, J$$

به‌طور مشابه می‌توان نشان داد که $\boldsymbol{\beta}_t$ از فرمول بازگشتی رو به عقب با شروع از زمان $t = T$ به‌صورت زیر محاسبه می‌شود

$$\boldsymbol{\beta}_T = \mathbf{1}, \ \mathbf{1} \in \mathbb{R}^J; \quad \boldsymbol{\beta}_t^\top = \boldsymbol{\Gamma}\mathbf{P}(y_{t+1})\boldsymbol{\beta}_{t+1}^\top, \quad t = 1, \ldots, T-1,$$

و تابع درستنمایی به فرم زیر به‌دست می‌آید:

$$\mathcal{L} = \Pr(y_1, \ldots, y_T) = \sum_{j=1}^{J} \alpha_t(j)\beta_t(j) = \boldsymbol{\alpha}_t^\top \boldsymbol{\beta}_t.$$

تابع درستنمایی مشاهده‌ها به‌صورت زیر است:

$$\mathcal{L} = \sum_{s_1, \ldots, s_T} \delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t} \prod_{t=1}^{T} f_{s_t}(y_t),$$

که با توجه به شکل جمعی و پیچیده آن، محاسبه برآورد بیشینه درستنمایی بر اساس آن بسیار پیچیده است. بنابراین برای به‌دست آوردن برآورد بیشینه درستنمایی از الگوریتم EM به‌صورت زیر استفاده می‌شود:

توابع نشانگر زیر را تعریف می‌کنیم

$$u_j(t) = \mathbf{I}(S_t = j), \quad v_{ij}(t) = \mathbf{I}(S_{t-1} = i, S_t = j); \quad t = 1, \ldots, T, \quad i, j = 1, \ldots, J \tag{۴}$$

تابع لگاریتم درستنمایی داده کامل[۱۳] یعنی تابع لگاریتم درستنمایی توأم مشاهده‌ها و وضعیت‌ها به‌صورت زیر است:

$$\begin{aligned}
\ell_c = \log \mathcal{L}_c &= \log \left( \delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t} \prod_{t=1}^{T} f_{s_t}(y_t) \right) \\
&= \log \delta s_t + \sum_{t=2}^{T} \log \gamma_{s_{t-1}, s_t} + \sum_{t=1}^{T} \log f_{s_t}(y_t) \\
&= \sum_{i=1}^{J} u_i(t) \log \delta_i + \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{t=2}^{T} v_{ij}(t) \log \gamma_{ij} + \sum_{i=1}^{J} \sum_{t=1}^{T} u_i(t) \log f_i(y_t).
\end{aligned}$$

بنابراین محاسبه امیدریاضی شرطی $\ell_c$ نسبت به‌توزیع $S_1, \ldots, S_T$ به‌شرط $y_1, \ldots, y_T$، مستلزم محاسبه امیدریاضی شرطی $u_j(t)$ها و $v_{ij}(t)$ها به شرط $y_1, \ldots, y_T$ و پارامترهای تکرار قبل به‌صورت زیر است.

[13] complete-data log-liklihood

با توجه به گام‌های پیشرو و پسرو داریم،

$$\begin{aligned}\hat{u}_j(t) &= \mathbb{E}(u_j(t)|y_1,\ldots,y_T,\boldsymbol{\theta}^{(r-1)}) \\ &= \Pr(S_t=j|y_1,\ldots,y_T) \\ &= \frac{\Pr(y_1,\ldots,y_t,S_t=j)\Pr(y_{t+1},\ldots,y_T|S_t=j)}{\sum_{j=1}^{J}\Pr(y_1,\ldots,y_T,S_t=j)} \\ &= \frac{\alpha_t(j)\beta_t(j)}{\boldsymbol{\alpha}_t^\top\boldsymbol{\beta}_t}; \quad t=1,\ldots,T, \quad j=1,\ldots,J.\end{aligned}$$

همچنین می‌توان نشان داد

$$\begin{aligned}\hat{v}_{ij}(t) &= \mathbb{E}(v_{ij}(t)|y_1,\ldots,y_T,\boldsymbol{\theta}^{(r-1)}) \\ &= \Pr(S_{t-1}=i,S_t=j|y_1,\ldots,y_T) \\ &= \frac{\Pr(y_1,\ldots,y_{t-1},S_{t-1}=i)\Pr(S_t=j|S_{t-1}=i)}{\sum_{j=1}^{J}\Pr(y_1,\ldots,y_T,S_t=j)} \\ &\quad\times\Pr(y_t|S_t=j)\Pr(y_{t+1},\ldots,y_T|S_t=j) \\ &= \frac{\alpha_{t-1}(i)\gamma_{ij}f_j(y_t)\beta_t(j)}{\boldsymbol{\alpha}_t^\top\boldsymbol{\beta}_t^\top}; \quad t=1,\ldots,T, \quad i,j=1,\ldots,J.\end{aligned}$$

گام M، شامل بیشینه‌سازی امیدریاضی شرطی تابع لگاریتم درستنمایی داده کامل، $\mathbb{E}\left(\ell_c|y_1,\ldots,y_T,\boldsymbol{\theta}^{(r-1)}\right)$ نسبت به پارامترها است، که به نتیجه‌های زیر می‌انجامد:

(۱) بیشینه‌سازی نسبت به $\delta_j$ منجر به برآوردی به فرم بسته زیر می‌شود

$$\hat{\delta}_j = \hat{u}_j(1); \quad j=1,\ldots,J$$

(۲) بیشینه‌سازی نسبت به $\gamma_{ij}$ منجر به برآوردی به فرم بسته زیر می‌شود

$$\hat{\gamma}_{ij} = \frac{\sum_{t=2}^{T}\hat{v}_{ij}(t)}{\sum_{k=1}^{J}\sum_{t=2}^{T}\hat{v}_{kj}(t)}; \quad i,j=1,\ldots,J.$$

(۳) برآورد پارامترهای توزیع $f$ که بسته به توزیع گسیل می‌تواند فرم بسته داشته باشد یا با استفاده از روش‌های عددی به‌دست آید.

برای برآورد هموار تابع‌های رگرسیونی برای $j=1,\ldots,J$ و $l=1,\ldots,p$، $f_l^{(j)}$ را به‌صورت بسط پایه‌های B-اسپلاین زیر مشخص می‌کنیم

$$f_l^{(j)}(x) = \sum_{k=1}^{K}\omega_{jlk}B_k(x),$$


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

که در آن $B_1, \ldots, B_K$ پایه‌های B-اسپلاین هستند. تعداد پایه‌های B-اسپلاین‌های درنظر گرفته شده میزان انعطاف‌پذیری تابع را مشخص می‌کند. رویکرد آیلر و مارکس [۱۰] و تاوانیدن تابع لگاریتم درستنمایی با اضافه کردن بخش تاوان تفاضلی مرتبه دوم روی ضریب‌های بسط B-اسپلاین را دنبال می‌کنیم. در این رویکرد تعداد پایه‌ها باید به اندازه کافی زیاد باشد، تا منجر به برآوردی انعطاف‌پذیر شود. درنظر گرفتن تفاضل مرتبه دوم که تقریبی از تاوان مربع انتگرال مشتق دوم است، منجر به تابع لگاریتم درستنمایی به فرم زیر می‌شود:

$$\ell_{c,p} = \ell_c - \sum_{j=1}^{J}\sum_{l=1}^{p}\frac{\lambda_{jl}}{2}\sum_{k=3}^{K}(\Delta^2\omega_{jlk})^2,$$

که در آن برای $j = 1, \ldots, J$ و $l = 1, \ldots, p$، $\lambda_{jl} > 0$ پارامترهای تاوان یا هموار هستند و $\Delta^2\omega_{jlk} = \omega_{jlk} - 2\omega_{jl,k-1} + \omega_{jl,k-2}$ است.

استفاده از تابع تاوان تفاضل مرتبه دوم اثر مشابهی با استفاده از تاوان نرم مشتق دوم دارد. زمانی‌که $\lambda_{jl} \longrightarrow \infty$، ضریب‌های مربوطه $\omega_{jl1}, \ldots, \omega_{jpK}$ به‌گونه‌ای برآورد می‌شوند، که بر روی یك خط صاف قرار گیرند. در نتیجه، به‌عنوان یك حالت خاص مدل رگرسیون خطی به‌دست می‌آید. برای دستیابی به توابع رگرسیون چندجمله‌ای مرتبه $q$ در حالت $\lambda_{jl} \longrightarrow \infty$ تفاضلات مرتبه $q+1$ در بخش تاوان استفاده کرد. در نتیجه مدل‌های رگرسیون خطی زیرمجموعه‌ای از برآورد جمعی تاوانیده هستند. همچنین مدل‌های جمعی استاندارد نیز با در نظر گرفتن $J = 1$، حالت خاصی مدل‌های جمعی تبدیل مارکوفی هستند.

برای اجرای الگوریتم EM تابع لگاریتم درستنمایی تاوانیده داده کامل را به فرم زیر بازنویسی می‌کنیم:

$$\begin{aligned}\ell_{c,p} = &\sum_{j=1}^{J} u_j(1)\log(\delta_j) + \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{t=2}^{T} v_{ij}(t)\log(\gamma_{ij}) \\ &+ \sum_{j=1}^{J}\sum_{t=1}^{T} u_j(t)\log\mathcal{N}\left(y_t; \eta^{(j)}(\mathbf{x_t}), \sigma^{(j)}\right) \\ &- \sum_{j=1}^{J}\sum_{p=1}^{P}\frac{\lambda_{jp}}{2}\sum_{k=3}^{K}\left(\Delta^2\omega_{jpk}\right)^2.\end{aligned}$$

گام E شامل محاسبه امیدهای شرطی $u_j(t)$ و $v_{ij}(t)$ به شرط داده‌ها و مقادیر برآوردشده در مرحله قبل هستند، که پیش‌تر بیان شدند.

گام M که شامل بیشینه‌سازی تابع درستنمایی داده کامل نسبت به پارامترهاست، برای برآورد پارامترهای $\delta_j$ و $\gamma_{ij}$ برای $i, j = 1, \ldots, J$ بیان شد. قسمت سوم یعنی بیشینه‌سازی تابع درستنمایی داده‌کامل نسبت به ضریب‌های $\omega_{jlk}$ معادل برآوردهای لگاریتم درستنمایی وزنی تاوانیده در مدل رگرسیونی استاندارد است، که برای هر وضعیت پنهان $j$ سهم مشاهده $t$ام در تابع لگاریتم درستنمایی دارای وزن $u_j(t)$ است.

با استفاده از رابطه زیر که جزئیات به‌دست آمدن آن در [۱۳] آمده است، می‌توانیم مقدار بهینه $\lambda_j$ها را محاسبه کنیم،

$$\lambda_j = \frac{K}{\hat{\boldsymbol{\beta}}\boldsymbol{\Omega}_{\boldsymbol{\lambda}_0}\hat{\boldsymbol{\beta}} + \mathrm{tr}(F\boldsymbol{\Omega}_{\boldsymbol{\lambda}_0})/n}, \quad j = 1, \ldots, d, \tag{۵}$$

که در $\boldsymbol{\lambda}_0$ مقدار آغازین بردار پارامترهای $\lambda_1, \ldots, \lambda_d$ است و

$$F = n(\mathbf{X}^\top\mathbf{X})^{-1}.$$

## ۴. رمزگشایی وضعیت‌های پنهان با الگوریتم ویتربی

تعریف می‌کنیم

$$\xi_{۱j} = \Pr(S_۱ = j, y_۱) = \delta_j f_j(y_۱)$$

و برای $t = ۲, \ldots, T$،

$$\xi_{tj} = \max_{s_۱,\ldots,s_{t-۱}} \Pr(S_۱ = s_۱, \ldots, S_{t-۱} = s_{t-۱}, S_t = j, Y_۱ = y_۱, \ldots, Y_t = y_t).$$

می‌توان نشان داد [۲۷] که برای $t = ۲, \ldots, T$ و $j = ۱, \ldots, J$ احتمال‌های بیشینه $\xi_{tj}$ در رابطه بازگشتی زیر صدق می‌کنند:

$$\xi_{tj} = \max_{۱\le i\le J} (\xi_{t-۱,i}\gamma_{ij})\, f_j(x_t).$$

محتمل‌ترین دنباله وضعیت‌ها را می‌توان از رابطه زیر به‌دست آورد

$$j_T^* = \arg\max_{j=۱,\ldots,J} \xi_{Tj},$$

و برای $t = T-۱, \ldots, ۱$

$$j_t^* = \arg\max_{j=۱,\ldots,J} \xi_{tj}\gamma_{j,j+۱}.$$

## ۵. پیش‌بینی وضعیت‌های آینده

برای پیش‌بینی وضعیت‌های آینده کافی‌است توزیع شرطی وضعیت‌های پنهان به شرط مشاهده‌ها را برای زمان $t > T$، بنویسیم. بدین ترتیب پیش‌بینی محتمل‌ترین وضعیت در هر زمان همانند آنچه در رمزگشایی موضعی بحث شد به‌دست می‌آید. تابع توزیع شرطی وضعیت به شرط مشاهده‌ها برای زمان $t > T$ به‌صورت زیر است [۲۷]

$$\Pr(S_t = j | Y_۱ = y_۱, \ldots, Y_T = y_T) = \frac{\boldsymbol{\alpha}_T \boldsymbol{\Gamma}^{t-T} \mathbf{e}_j^\top}{\boldsymbol{\alpha}_t^\top \boldsymbol{\beta}_t},$$

که در آن $\mathbf{e}_j^\top$ یك بردار به طول $J$ است که به‌جز درایه $j$ام آن که برابر با یك است، مابقی برابر با صفر هستند. $\boldsymbol{\Gamma}^{t-T}$ توان $t-T$ ماتریس $\boldsymbol{\Gamma}$ است، که از $t-T$ مرتبه ضرب این ماتریس در خودش به‌دست می‌آید.

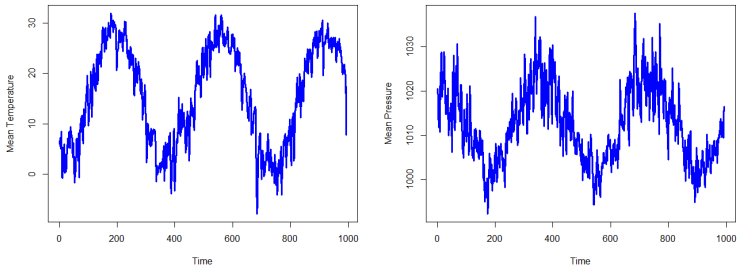
## ۶. بسته نرم‌افزاری hhsmm در R

بسته نرم‌افزاری hhsmm در نرم‌افزار R شامل تمامی ابزارهای لازم برای آغاز، مدل‌بندی، برازش و پیش‌بینی به کمک مدل‌های مارکوف، نیمه مارکوف، ترکیبی از مارکوف و نیمه مارکوف، مدل رگرسیون تغییر رژیم مارکوفی خطی پارامتری، ناپارامتری جمعی و انتقالی است. این بسته نرم‌افزاری در سایت CRAN به آدرس https://cran.r-project.org/package=hhsmm قرار دارد و پیوسته مورد ارزیابی و نگهداری قرار گرفته است و تاکنون نسخه‌های مختلفی از آن به روزرسانی شده است. برای اطلاعات بیشتر در خصوص این بسته نرم‌افزاری می‌توانید علاوه بر مطالعه راهنمای بسته، مقاله [۳] را نیز مطالعه کنید.


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

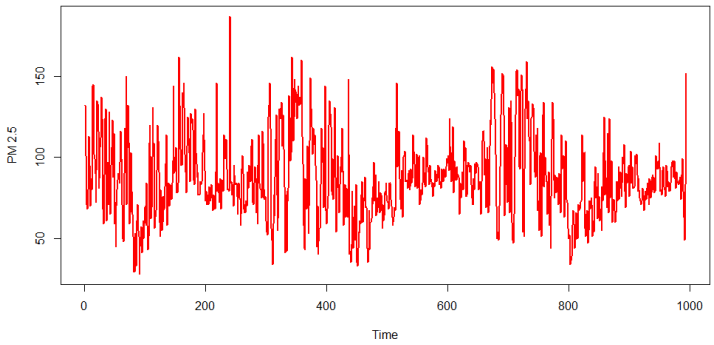
## ۷. تحلیل داده‌های ذرات معلق در هوای شهر تهران

داده‌های این مقاله از دو منبع مختلف به‌دست آمده و همگام‌سازی شده‌اند. داده‌های مربوط به ذرات معلق PM۲.۵ که از سایت https://airnow.tehran.ir/ به دست آمده‌اند و همچنین داده‌های مربوط به آب و هوا که از سازمان هواشناسی کشور اخذ شده‌اند. این داده‌ها پس از همگام سازی شامل ۹۹۳ روز از ابتدای سال میلادی ۲۰۱۵ تا آخر سپتامبر ۲۰۱۷ هستند. از این داده‌ها سه متغیر دمای هوا، فشار هوا و میزان ذرات معلق PM۲.۵، برای این تحلیل در نظر گرفته شده‌اند. مقدار ذرات PM۲.۵، میانگین تمامی ایستگاه‌های شهر تهران است. همچنین مقدارهای دما و فشار هوا نیز برای تمامی ساعات روزها و تمامی ایستگاه‌های تهران و حومه میانگین گرفته شده‌اند. نمودار شکل ۳، دما و فشار هوای شهر تهران را در این بازه زمانی نشان می‌دهد. همچنین نمودار سری زمانی ذرات PM۲.۵ در نمودار شکل ۴ آورده شده است.




**شکل ۳.** دما (چپ) و فشار (راست) هوای شهر تهران در فاصله سال‌های ۱۳۹۴ تا ۱۳۹۶

**Figure** 3: Temperature (left) and pressure (right) of Tehran city air during 2015-2017

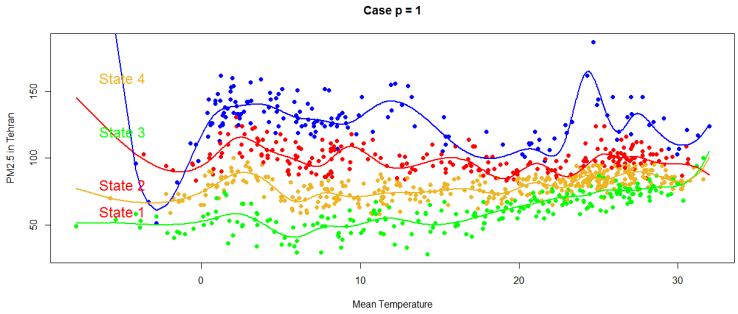



**شکل ۴.** ذرات معلق PM۲.۵ در هوای شهر تهران در فاصله سال‌های ۱۳۹۴ تا ۱۳۹۶

**Figure** 4: PM2.5 in Tehran city air during 2015-2017

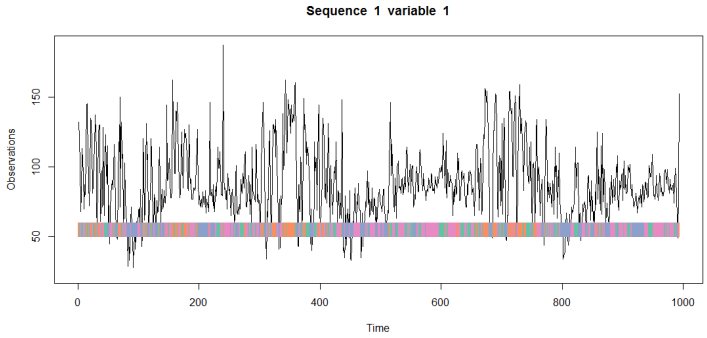
برای درک نحوه برازش مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی، ابتدا تنها دمای هوا را به‌عنوان متغیر پیش‌گو و میزان ذرات PM۲.۵ را به‌عنوان متغیر پاسخ، در نظر می‌گیریم. با توجه به ساختار فصلی داده‌ها و تغییرات چهار فصل سال، تعداد وضعیت‌های پنهان را برابر ۴ قرار می‌دهیم.

حال مدل (۲) را با $p = ۱$ و $J = ۴$ به داده‌ها برازش می‌دهیم. نمودار ۵ پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی تنها بر اساس دمای هوا در هر یک از چهار وضعیت پنهان را نشان می‌دهد. این نمودار به زیبایی چهار خوشه (وضعیت پنهان) و چهار مدل رگرسیون ناخطی ناپارامتری که بر اساس اسپلاین‌های B-هموار تاوانیده به نقاط هر خوشه برازش داده شده نشان می‌دهد.




**شکل ۵.** پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی تنها بر اساس دمای هوا در هر یک از چهار وضعیت پنهان. نقاط هر وضعیت با رنگ مختلف تفکیک شده و نمودار مقدارهای پیش‌بینی شده با همان رنگ رسم شده است.

**Figure** 5: 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.



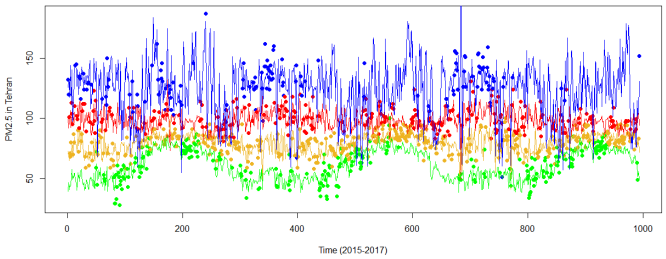

**شکل ۶.** پیش‌بینی وضعیت‌های پنهان دنباله ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی (رنگ‌های نوار پایین نمودار)

**Figure** 6: Prediction of the hidden states for the sequence of PM2.5 in Tehran city air using the fitted regime switching regression model (the colors of the bar at the bottom of the plot).

حال متغیر پیشگوی فشار هوا را نیز به مدل اضافه کرده و حالت $p = ۲$ و $J = ۴$ مدل (۲) را به داده‌ها برازش می‌دهیم. نمودار شکل ۶ وضعیت‌های پنهان پیش‌بینی شده (نتیجه خوشه‌بندی الگوریتم) را که با استفاده از هموارسازی (یعنی بیشینه‌سازی احتمال‌های الگوریتم پیش‌رو–پس‌رو و بدون استفاده از الگوریتم ویتربی) به‌دست آمده‌اند، به‌صورت چهار رنگ مختلف در نواری در زیر نمودار سری زمانی میزان PM۲.۵ نشان می‌دهد.

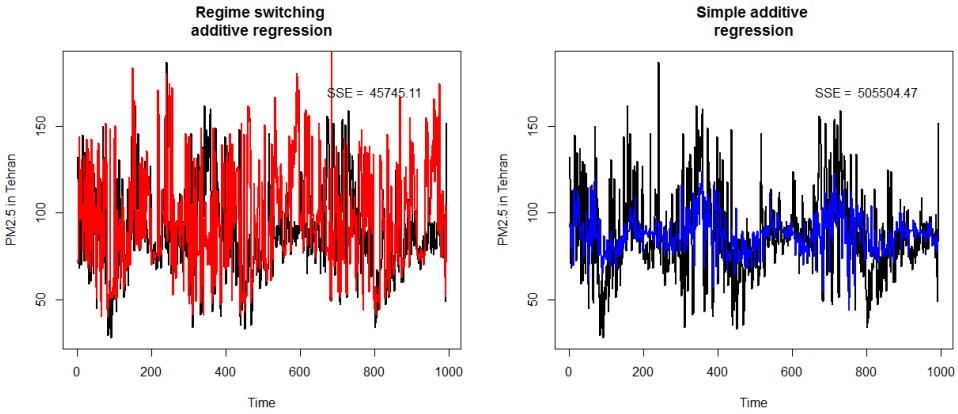
برای پیش‌بینی مقدار PM۲.۵ از روی متغیرهای پیشگوی دما و فشار، از پارامترهای ضریب رگرسیونی برآورد شده برای هر وضعیت پنهان مانند یک مدل رگرسیون جمعی عادی (تک وضعیته) استفاده می‌کنیم، با این تفاوت که در ابتدا این پیش‌بینی‌ها به‌صورت مجزا برای هر وضعیت پنهان به‌دست می‌آیند و برای به‌دست آوردن پیش‌بینی نهایی بایستی آن‌ها را با هم ترکیب کرد. نمودار شکل ۷، این پیش‌بینی‌های مجزا را با چهار رنگ مختلف برای چهار وضعیت نشان می‌دهد. همچنین نقاطی از داده که توسط روش هموارسازی به چهار خوشه تخصیص داده شده‌اند نیز با همان رنگ‌ها تفکیک شده‌اند. در این نمودار می‌توان به خوبی تطبیق پیش‌بینی‌های هر خوشه را با مشاهدات همان خوشه، مشاهده کرد.


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




**شکل ۷.** پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی بر اساس دما و فشار هوا در هر یک از چهار وضعیت پنهان. نقاط هر وضعیت با رنگ مختلف تفکیک شده و نمودار مقدارهای پیش‌بینی شده با همان رنگ رسم شده است.

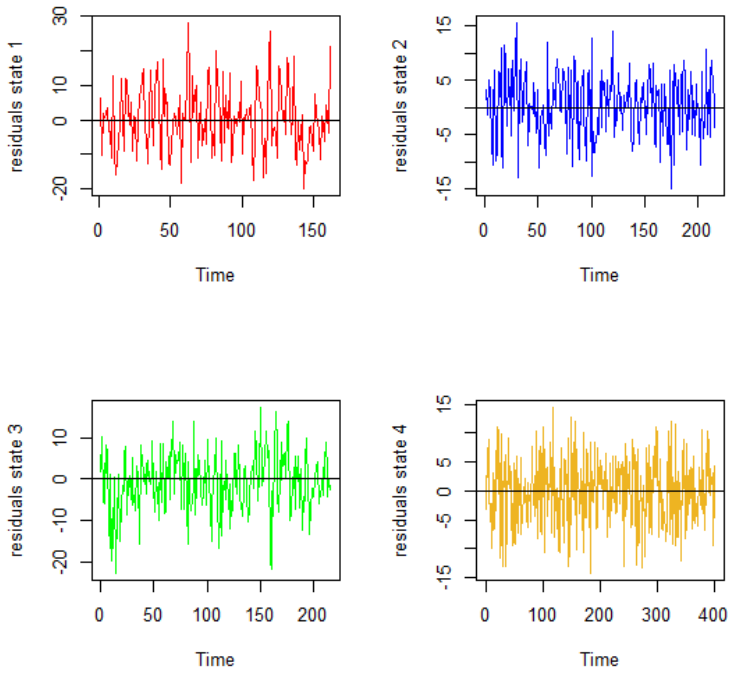
**Figure** 7: Prediction of PM2.5 in Tehran city air using regime switching regression model, based on the air temperature and pressure, in each of the four hidden states. The points in each state are colored by different colors and the curve of the prediction is drawn with the same color.




**شکل ۸.** مقایسه دقت پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط دو مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی با چهار وضعیت پنهان و مدل رگرسیون ناپارامتری جمعی تک وضعیته. میزان مجموع توان‌های دوم خطای هر یک از مدل‌ها در نمودار مشخص شده است.

**Figure** 8: 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.

برای ترکیب این پیش‌بینی‌های مجزا با یکدیگر و به‌دست آوردن پیش‌بینی نهایی کافی است مقدارهای پیش‌بینی شده را به ترتیب مشاهده‌های سری زمانی در نظر گرفت، یعنی پیش‌بینی نهایی برای هر مشاهده برابر است با پیش‌بینی به‌دست آمده برای خوشه (وضعیت پنهان) متناظر با آن مشاهده. بدین ترتیب به نمودار شکل ۸ (چپ) می‌رسیم. در این شکل خطوط سیاه مقدارهای واقعی PM۲.۵ و خطوط قرمز مقدارهای پیش‌بینی شده هستند. همچین مقدار مجموع توان دوم خطای برازش نیز $SSE = ۴۵۷۴۵/۱۱$ به‌دست آمده است و در نمودار نشان داده شده است.




**شکل ۹.** نمودارهای مانده در مقابل زمان به تفکیک وضعیت پنهان مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی

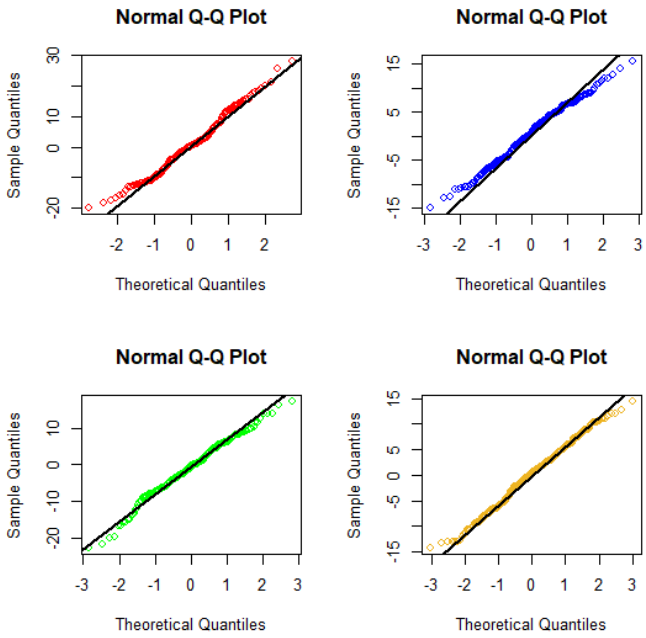
**Figure** 9: The plots of residuals versus time for each hidden state of the regime switching nonparametric additive regression model.

برای بررسی عملکرد مدل ارائه شده، این مدل را با مدل رگرسیون ناپارامتری جمعی (تک وضعیته) مقایسه کرده‌ایم. بدین منظور، مدل رقیب را نیز به داده‌ها برازش داده و مقدارهای پیش‌بینی شده را به همراه مقدار واقعی دنباله PM۲.۵ در نمودار شکل ۸ (راست) رسم کرده‌ایم. در این نمودار نیز مقدارهای واقعی PM۲.۵ با خط‌های سیاه و مقدارهای پیش‌بینی شده با خط‌های آبی نشان داده شده است. همچین مقدار مجموع توان دوم خطای برازش نیز $SSE = ۵۰۵۵۰۴/۴۷$ به‌دست آمده است و در نمودار نشان داده شده است، که کارایی ۱۱ برابری روش رگرسیون تغییر رژیم مارکوفی با چهار وضعیت پنهان را نسبت به مدل رگرسیون ناپارامتری جمعی تک وضعیته نشان می‌دهد.

با توجه به مثبت بودن مقدار PM۲.۵ تحلیل مانده‌های مدل می‌تواند به‌بررسی فرض نرمال بودن خطاهای مدل رگرسیون تغییر رژیم مارکوفی کمک کند. برای این کار مانده‌ها را در هر وضعیت به‌دست آورده‌ایم و نمودار مانده‌ها در مقابل زمان را در شکل ۹ و نمودار چندک‌های نرمال Q–Q مانده‌ها را در شکل ۱۰ آورده‌ایم. اگرچه این نمودارها انحراف‌های اندکی را از فرض نرمال بودن نشان می‌دهند ولی انحراف‌ها قابل ملاحظه نیستند و می‌توان از این انحراف اندک چشم پوشی کرد. با این حال انجام برخی تبدیل‌ها مانند تبدیل لگاریتمی بر روی PM۲.۵ و همچنین متغیرهای دما و فشار می‌تواند این انحراف‌های اندک را نیز کاهش دهد.

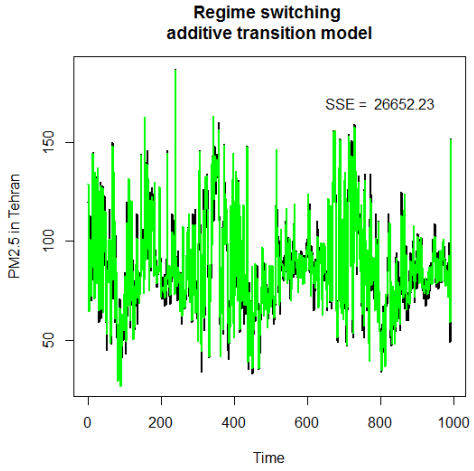
با این‌حال هیچ کدام از مدل‌های بیان شده در بالا نمی‌توانند برای پیش‌بینی زمان آینده متغیر پاسخ، مفید باشند. برای پیش‌بینی زمان آینده PM۲.۵ از روی مقدارهای زمان حال دما و فشار هوا بایستی مدل رگرسیون انتقالی تغییر رژیم مارکوفی را به داده‌ها برازش دهیم. ابتدا تأخیر $\ell = L = ۱$ را در نظر می‌گیریم (تأخیر یک روزه). نمودار شکل ۱۱ نتیجه پیش‌بینی این مدل و مجموع توان‌های دوم خطای آن را برابر $SSE = ۲۶۶۵۲/۲۳$ نشان می‌دهد، که نسبت به مدل بدون تأخیر دارای کارایی حدودا ۱/۷ برابری است.


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



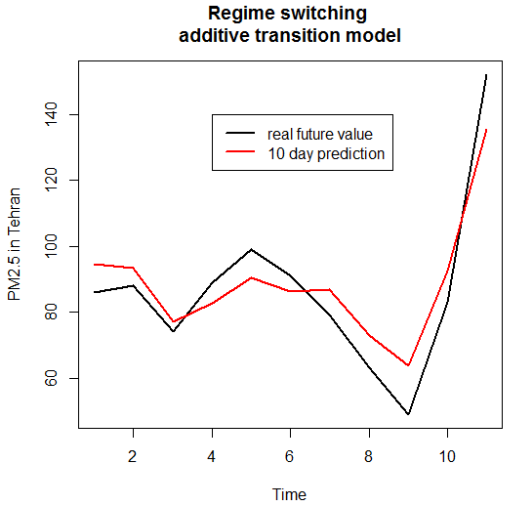

**شکل ۱۰.** نمودارهای نمودار چندک‌های نرمال Q–Q به تفکیک وضعیت پنهان مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی

**Figure** 10: The Q-Q plots of normal quantiles for each hidden state of the regime switching nonparametric additive regression model.




**شکل ۱۱.** پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی انتقالی با تأخیر یک روزه. میزان مجموع توان‌های دوم خطای مدل در نمودار مشخص شده است.

**Figure** 11: Prediction of PM2.5 in Tehran city air using transitive regime switching nonparametric additive regression model with 1 day lag. The mean squared error of the model is presented in the plot.




**شکل ۱۲.** پیش‌بینی خارج از نمونه ده روزه میزان ذرات معلق PM۲.۵ در هوای شهر تهران توسط مدل برازش داده شده رگرسیون تغییر رژیم مارکوفی انتقالی با تأخیر ۱۰ روزه. میزان میانگین توان‌های دوم پیش‌بینی مدل برابر ۹۹/۳ است.

**Figure** 12: Out of sample prediction of PM2.5 in Tehran city air using transitive regime switching nonparametric additive regression model with 10 days lag. The mean squared error of the model is equal to 99.3.

با این‌حال پیش‌بینی به‌دست آمده برای داده‌های آموزشی است و آن‌چه می‌تواند کارایی واقعی این مدل را نشان دهد، پیش‌بینی واقعی آینده، یعنی پیش‌بینی داده‌های خارج از مجموعه داده آموزشی است. برای این کار یک مدل با تأخیر $\ell = L = ۱۰$ در نظر می‌گیریم (تأخیر ده روزه) و آن را با استفاده از تمامی روزها به‌جز ۱۰ روز آخر آموزش داده‌ایم. سپس از این مدل برای پیش‌بینی ده روز حذف شده از انتهای داده‌ها، استفاده کرده‌ایم. نمودار شکل ۱۲ نتیجه را نشان می‌دهد. مقدار میانگین توان‌های دوم خطای پیش‌بینی برای این ده روز برابر $MSPE = ۹۹/۳$ به‌دست آمده است. خط سیاه در این نمودار مقدارهای واقعی PM۲.۵ در ده روز و خط‌های قرمز مقدارهای پیش‌بینی شده را نشان می‌دهد. همان‌طور که از نمودار می‌توان دید، پیش‌بینی انجام شده بسیار مطلوب است. بایستی این مسأله را در نظر گرفت که در مسائل آماری، پیش‌بینی زمان آینده در بازه‌های طولانی برای چنین داده‌ها با ساختار پیچیده‌ای آن هم بر اساس تعداد کم متغیرهای پیشگو کار دشواری است و دستیابی به چنین دقتی بسیار با اهمیت است. علاوه بر این در این پژوهش تنها از دما و فشار هوا استفاده شده است و متغیرهایی مانند سرعت وزش باد و رطوبت هوا که می‌توانند به افزایش دقت پیش‌بینی میزان آلاینده‌هایی مانند PM۲.۵ کمک بسیاری کنند به‌کار برده نشده‌اند. این مسائل به خوبی اهمیت دستیابی به چنین نتایج دقیقی را به‌خصوص برای داده‌های خارج از مجموعه داده‌های آموزشی، روشن می‌سازند.

## ۸. جمع‌بندی و نتیجه‌گیری

در این مقاله به‌معرفی مدل رگرسیون تغییر رژیم مارکوفی ناپارامتری جمعی و حالت خاص آن، که شامل تأخیرهای زمانی متغیر پاسخ و متغیرهای کمی است، پرداختیم. از این مدل برای پیش‌بینی میزان ذرات معلق PM۲.۵ در هوای تهران با کمک متغیرهای پیش‌گوی دما و فشار هوا، بهره گرفتیم. نتایج به‌دست آمده نشان دادند که مدل رگرسیون تغییر رژیم مارکوفی، که به نوعی یک مدل رگرسیون خوشه‌بندی شده با یک مدل مارکوف پنهان است، بسیار بهتر از مدل رگرسیون ناپارامتری جمعی عمل می‌کند. همچنین حالت انتقالی این مدل می‌تواند برای پیش‌بینی مقدارهای آینده متغیر پاسخ، با اطمینان خوبی به‌کار بسته شود.


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

تمامی ابزارهای لازم برای مدل‌بندی، برازش و پیش‌بینی به کمک این مدل‌ها در بسته نرم‌افزاری hhsmm در نرم افزار R توسعه داده شده‌اند.

از برخی پیشنهادات برای تحقیقات آینده که توسط نویسنده نیز در دست انجام است می‌توان به تحلیل توأم داده‌های آلودگی هوای تهران، تحلیل فضایی–زمانی بر اساس مدل‌های ارائه شده، انتخاب متغیر در مدل‌های تغییر رژیم مارکوفی و ارائه مدل‌های مقاوم در برابر مشاهده‌های دورافتاده، اشاره کرد.

## مراجع


[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.

**سید مرتضی امینی**
بخش آمار، دانشکده ریاضی آمار و علوم کامپیوتر، دانشکدگان علوم، دانشگاه تهران
morteza.amini@ut.ac.ir

سید مرتضی امینی متولد شهریور ماه ۱۳۶۰ در شهر مشهد است. وی در سال ۱۳۷۸ وارد مقطع کارشناسی رشته آمار دانشگاه شهید باهنر کرمان شد و در سال ۱۳۸۲ مدرک کارشناسی آمار خود را دریافت نمود. سپس در فاصله سال‌های ۱۳۸۳ تا ۱۳۸۶ به تحصیل در دوره کارشناسی ارشد آمار–ریاضی در دانشگاه فردوسی مشهد پرداخت. او در سال ۱۳۸۶ در دوره دکتری تخصصی آمار در دانشگاه فردوسی مشهد پذیرفته شد و در سال ۱۳۹۰ توانست مدرک دکتری خود را دریافت نماید. از سال ۱۳۹۰ در دانشگاه تهران به عنوان عضو هیات علمی مشغول به کار شد. در حال حاضر او دانشیار آمار دانشگاه تهران است.


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