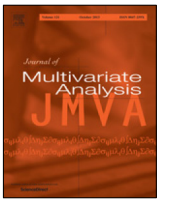
Journal of Multivariate Analysis 179 (2020) 104646

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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



Continuous time hidden Markov model for longitudinal data

Jie Zhou [a], Xinyuan Song [b,*], Liuquan Sun [c]

[a] School of Mathematics, Capital Normal University, Beijing, 100048, PR China
[b] Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[c] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, PR China

ARTICLE INFO

Article history:
Received 22 October 2019
Received in revised form 30 May 2020
Accepted 30 May 2020
Available online 10 June 2020

AMS 2010 subject classifications:
05B05
05B25

Keywords:
Continuous-time HMMs
Longitudinal data
ML estimator
Unknown number of hidden states
SCAD penalty

ABSTRACT

Hidden Markov models (HMMs) describe the relationship between two stochastic processes, namely, an observed outcome process and an unobservable finite-state transition process. Given their ability to model dynamic heterogeneity, HMMs are extensively used to analyze heterogeneous longitudinal data. A majority of early developments in HMMs assume that observation times are discrete and regular. This assumption is often unrealistic in substantive research settings where subjects are intermittently seen and the observation times are continuous or not predetermined. However, available works in this direction restricted only to certain special cases with a homogeneous generating matrix for the Markov process. Moreover, early developments have mainly assumed that the number of hidden states of an HMM is fixed and predetermined based on the knowledge of the subjects or a certain criterion. In this article, we consider a general continuous-time HMM with a covariate specific generating matrix and an unknown number of hidden states. The proposed model is highly flexible, thereby enabling it to accommodate different types of longitudinal data that are regularly, irregularly, or continuously collected. We develop a maximum likelihood approach along with an efficient computer algorithm for parameter estimation. We propose a new penalized procedure to select the number of hidden states. The asymptotic properties of the estimators of the parameters and number of hidden states are established. Various satisfactory features, including the finite sample performance of the proposed methodology, are demonstrated through simulation studies. The application of the proposed model to a dataset of bladder tumors is presented.

1. Introduction

Longitudinal data often arise in empirical investigations when subjects are monitored over a period of time. Hidden Markov models (HMMs) are well suited for characterizing longitudinal data in terms of a set of states because they describe the relationship between two stochastic processes, namely, an observable outcome process and an underlying hidden state process. Given their ability to simultaneously reveal the dynamic association and heterogeneity of a longitudinal process, HMMs have attracted significant attention from various disciplines [1,2,6,9,18,19,28,32,34,40]. However, the preceding works have focused on discrete-time HMMs, which assume longitudinal responses are observed at regular times $t \in \{1, \ldots, T\}$. In many instances during clinical trials and observational studies, observation times are irregular or

* Corresponding author.
E-mail address: xysong@sta.cuhk.edu.hk (X. Song).

https://doi.org/10.1016/j.jmva.2020.104646
0047-259X/

continuous. Thus, the state dwell-time is unlikely to be appropriately modeled through such an implicit state occupancy distribution. Despite several early developments in continuous-time Markov process [4,31], available methods in this direction are still limited and under development. Guédon [15] and Langrock and Zucchini [21] proposed a hidden semi-Markov model, which formulated the state dwell-time through a positive integer random variable and separately modeled the transition probability matrix of the embedded Markov chain, to allow irregular observation times. However, these approaches only enabled the modeling of discrete sequences with time going to infinity and did not investigate the asymptotic properties of parameter estimators. For a situation with a continuous observation time, Guihenneuc-Jouyaux et al. [16] considered a simple case of a hidden Markov process with a known number of states and specific transition patterns. Fearnhead and Sherlock [13] and Scott and Smyth [33] proposed a Markov-modulated Poisson process to analyze click rate data and the occurrence of a rare DNA motif, respectively. Bureau et al. [8] considered a continuous-time HMM with only two states for longitudinal data. Liu et al. [25] studied efficient learning of a continuous-time HMM for disease progression. Their models allowed a large number of states to be defined roughly by some observed surrogate covariates, which may not be the case in practice. Moreover, their developments focused only on special cases in which either the outcome process was restricted to the Poisson process or the state process was assumed to have a special transition structure. Recently, Langrock et al. [20] conducted estimation for stochastic volatility models using structured HMMs and showed that discrete HMMs can be used to approximate their continuous counterpart. Bartolucci and Farcomeni [5] constructed a joint model for longitudinal data and survival time via a continuous-time hidden Markov process. Nevertheless, the number of hidden states in all above studies was assumed to be known a priori.

In substantive research, the number of hidden states of an HMM is usually unknown. Although information criterion-based methods, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), can be used for selecting the number of hidden states, their use in the context of continuous-time HMMs has not been theoretically justified [26]. MacKAY [27] developed a penalized minimum-distance approach to obtain a consistent estimator of the number of states for discrete-time HMMs. Chen and Khalili [10] extended this approach to a double penalized log-likelihood method for mixture models. Hung et al. [17] considered the use of the double penalized log-likelihood method for discrete-time HMMs. However, these methods assumed discrete times and restricted a special case where the states were determined by a one-dimensional parameter, and is thus inapplicable to the general case considered in the present study. Farcomeni [12] developed a penalized likelihood approach to obtain robust estimation for more general discrete-time HMMs, but the focus of penalization is on parameter estimation rather than on selection of the number of hidden states. To our knowledge, no existing method is available to the proposed context.

In this article, we propose a continuous-time HMM with an unknown number of states and transition patterns. The proposed model, in which the between-transition of states is described by a hidden Markov process instead of a Markov chain, is able to accommodate different types of longitudinal data that are regularly, irregularly, or continuously collected. We propose to directly maximize the observed-data log-likelihood function for estimation by using formula for differentiating and integrating exponential matrices and the related theorems of exponential matrix. Good asymptotic properties of the resulting maximum likelihood (ML) estimators are established. We also develop a new penalized likelihood method to estimate the number of hidden states of HMMs and show the consistency of the derived estimator. The main contribution of this study can be summarized as follows. First, the proposed model provides a general framework to manage longitudinal data that are regularly, irregularly, or continuously collected. Second, we apply differential theorems to directly maximize the observed-data log-likelihood function, thereby leading to an efficient estimation procedure. Finally, we propose a novel penalized method to estimate the number of hidden states of HMMs and establish the consistency of the resulting estimator. Despite of high importance, such a penalized procedure for selection of the number of hidden states has never been investigated in the context of continuous-time HMMs. The proposed penalization is not a simple application of traditional penalized methods because it takes into account the dynamic transition feature of HMMs.

The rest of the paper is organized as follows. Section 2 describes the proposed model. Section 3 presents an estimation procedure and the associated asymptotic properties. Section 4 proposes a novel penalized approach to select the number of hidden states and establishes the consistency of the resulting estimator. Section 5 conducts simulation studies to evaluate the proposed methods. Section 6 reports an application to the bladder cancer dataset. Section 7 concludes the paper. Technical details are provided in the appendix and an online supplementary material.

2. Model specification

Let $Y(t)$ be the value of longitudinal response at time t, $X(t)$ be a $p \times 1$ vector of covariates, and $S(t)$ be the underlying hidden state at time t and take values in a finite set $\{1, \ldots, d\}$. The longitudinal response and covariates are observed at discrete times, which are assumed to be independent of the hidden states and longitudinal responses. Assume that the longitudinal response variable $Y(t)$ has a probability density function $f(y; x, \beta_k, \phi)$ given covariate value $X(t) = x$ and the underlying state $S(t) = k$, $k \in \{1, \ldots, d\}$, where β_k is the $p \times 1$ regression parameter of state k and ϕ is an $r \times 1$ common parameter of all states. The hidden state $S(t)$ is assumed to be a continuous time Markov process with a generating matrix $\boldsymbol{G}$ that satisfies the following conditions:

$$\begin{cases} g_{kl} > 0, & k \neq l, \\ g_{kk} = -\sum_{l \neq k} g_{kl}, & k \in \{1, \ldots, d\}. \end{cases} \tag{1}$$

According to [31], the definition of the generating matrix implies (i) the duration of a visit to state k follows an exponential distribution with rate $-g_{kk}$, (ii) staying on state k, the transition probability from state k to state $l \neq k$ is $P_{kl} = -g_{kl}/g_{kk}$, and (iii) from time 0 to t, $S(t)$ has transition probability matrix $\boldsymbol{P}(t) = \exp\{\boldsymbol{G}t\}$. To allow the generating matrix to be subject specific, we assume that, for $k \neq j$, the (k, j)th element of $\boldsymbol{G}$ has the form of $g_{kj}(Z) = \exp\{\gamma_{kj}'Z\}$, where γ_{kj} is a $s \times 1$ vector of unknown parameters and Z is a $s \times 1$ vector of baseline covariates that may contain the time-independent part of X. Under this model, the generating matrix can depend on certain subject-specific characteristics.

Remark 1. For a discrete-time HMM with transition probability matrix $\boldsymbol{Q}$, we can embed a continuous-time HMM by taking the generating matrix as $\boldsymbol{G} = \ln(\boldsymbol{Q})$ [22], which corresponds to the model with $Z \equiv 1$. This logarithm does not always exist because $\boldsymbol{G}$ must be a real matrix. Gallier [14] provided a necessary and sufficient condition for a real matrix to possess a real logarithm. In particular, if all the eigenvalues of $\boldsymbol{Q}$ are positive, then $\boldsymbol{G}$ is well defined. Thus, the following proposed estimation procedure can be applied to the discrete time situation.

For the distribution of the initial state, let $\pi_k(W) = Pr(S(0) = k|W)$, where W is a $q \times 1$ vector of baseline covariates, which may contain the time-independent part of X. In addition, we assume that

$$\pi_k(W) = \frac{\exp\{\eta_k'W\}}{\sum_{l=1}^{d}\exp\{\eta_l'W\}}, \quad k \in \{1, \ldots, d\}, \tag{2}$$

where $\eta_1, \ldots, \eta_{d-1}$ are unknown parameters and η_d is a $q \times 1$ vector of zeros.

Liu et al. [25] also proposed continuous-time HMMs to examine disease progression and developed an efficient learning algorithm for estimation. However, their model was designed for sequences and did not consider subject-specific covariate effects. By contrast, the proposed model assumes that the longitudinal response follows a known distribution with density $f(\cdot)$, which is parametrized by a function of covariates and hidden state dependent "regression" coefficients. Moreover, the proposed model allows for subject-specific generating matrix and initial values, which elicit subject-specific hidden trajectories. Therefore, the proposed model can be regarded as a generalization of GLMMs because it allows for outcome-covariate association heterogeneity driven by transitions in Markovian hidden regimes.

3. Maximum likelihood estimator

For an independently and identically distributed random sample of n subjects, the observed data comprise $\{W_i, Z_i, t_{ij}, X_{ij}, Y_{ij}, i \in \{1, \ldots, n\}, j \in \{1, \ldots, n_i\}\}$, where t_{ij} denotes the jth observation time for subject i, and X_{ij} and Y_{ij} are the observations of $X(\cdot)$ and $Y(\cdot)$ at time t_{ij}. Define $\tilde{t}_{ij} = t_{ij} - t_{i,j-1}$ for $j \in \{1, \ldots, n_i\}$ with $t_{i0} = 0$, let $\tilde{\boldsymbol{Y}}_{ij}$ be the $d \times d$ diagonal matrix with the kth diagonal element being the density function $f(Y_{ij}; X_{ij}, \beta_k, \phi)$, $\boldsymbol{G}_i = (g_{ikl})$ be the generating matrix of the hidden Markov process of the ith subject, and denote $g_{ikl} = g_{kl}(Z_i)$. Then, the transition matrix from $t_{i,j-1}$ to t_{ij} is $\exp\{\boldsymbol{G}_i\tilde{t}_{ij}\}$. Therefore, similar to the arguments of [13], we derive that for subject i, the observed-data likelihood function when the chain ends in state l given that it starts in state k is

$$L(Y_{ij}, j \in \{1, \ldots n_i\}, S_i(t_{in_i}) = l|S_i(0) = k, X_i) = e_k' \prod_{j=1}^{n_i} \boldsymbol{F}_{ij} e_l, \tag{3}$$

where $\boldsymbol{F}_{ij} = \exp(\boldsymbol{G}_i\tilde{t}_{ij})\tilde{\boldsymbol{Y}}_{ij}$ and e_k is a column vector with the kth component being one and all other components being zero.

3.1. Estimation procedure and computer algorithm

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{d-1})$, and $\theta = (\text{vec}(\boldsymbol{\beta})', \text{vec}(\boldsymbol{\eta})', \gamma_{kj}', k \in \{1, \ldots, d\}, j \neq k, \phi')'$, where vec$(\cdot)$ denotes the operator that transforms a matrix into a vector by stacking the columns of the matrix one underneath the other. Then, we can write the observed-data likelihood function for the ith subject as follows:

$$L_i(\theta) = \sum_{k=1}^{d}\sum_{l=1}^{d} \pi_{ik} L(Y_{ij}, j \in \{1, \ldots n_i\}, S_i(t_{in_i}) = l|S_i(0) = k, X_i) = \pi_i' \prod_{j=1}^{n_i} \boldsymbol{F}_{ij}\boldsymbol{1}, \tag{4}$$

where $\pi_{ik} = \pi_k(W_i)$, $\pi_i = (\pi_{i1}, \ldots, \pi_{id})'$ and the $\boldsymbol{1}$ denotes the column vector with all components being 1. Then, the observed-data log-likelihood function is $L(\theta) = \sum_{i=1}^{n} \log L_i(\theta)$. The ML estimator of θ can be obtained through the maximization of $L(\theta)$. Although $L(\theta)$ has a simple form, it is mathematically complicated because it includes the exponential operation of a matrix. The derivatives of $L(\theta)$ with respect to γ can be obtained by utilizing the related theorems of exponential matrix. Recalling that for any matrix function $\boldsymbol{A}(t) = (a_{ij}(t))$, the first derivative of $\exp(\boldsymbol{A}(t))$ with respect to t has the following form [39]:

$$\frac{d}{dt}\exp\{\boldsymbol{A}(t)\} = \int_0^1 \exp\{u\boldsymbol{A}(t)\}\dot{\boldsymbol{A}}(t)\exp\{(1-u)\boldsymbol{A}(t)\}du, \tag{5}$$

where $\dot{\boldsymbol{A}}(t) = (\dot{a}_{ij}(t))$ and $\dot{a}_{ij}(t) = da_{ij}(t)/dt$. By using the formula of Theorem 1 in [38], we can calculate the integral (5). Specifically, for any $d \times d$ matrices $\boldsymbol{B}$ and $\boldsymbol{C}$,

$$\int_0^1 \exp\{u\boldsymbol{B}\}\boldsymbol{C}\exp\{(1-u)\boldsymbol{B}\}du = (\exp\{\boldsymbol{H}\})_{1:d,(d+1):(2d)}, \quad \boldsymbol{H} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}. \tag{6}$$

On the basis of (5), we obtain the partial derivatives of $L(\theta)$ with respect to the elements of θ as follows:

$$\begin{aligned}
\frac{\partial}{\partial \beta_{mk}} L_i(\theta) &= \sum_{j=1}^{n_i} \pi_i' \Big\{ \prod_{v=1}^{j-1} \boldsymbol{F}_{iv}(\theta) \Big\} \boldsymbol{D}_{\beta_{mk}}^{(ij)}(\theta) \Big\{ \prod_{v=j+1}^{n_i} \boldsymbol{F}_{iv}(\theta) \Big\} \mathbf{1}, \quad m \in \{1,\ldots,p\},\ k \in \{1,\ldots,d\}, \\
\frac{\partial}{\partial \eta_{mk}} L_i(\theta) &= W_{im}\pi_{ik}(e_k - \pi_i)' \prod_{j=1}^{n_i} \boldsymbol{F}_{ij}(\theta)\mathbf{1}, \quad m \in \{1,\ldots,q\},\ k \in \{1,\ldots,d-1\}, \\
\frac{\partial}{\partial \gamma_{klm}} L_i(\theta) &= \sum_{j=1}^{n_i} \pi_i' \Big\{ \prod_{v=1}^{j-1} \boldsymbol{F}_{iv}(\theta) \Big\} \boldsymbol{D}_{\gamma_{klm}}^{(ij)}(\theta) \Big\{ \prod_{v=j+1}^{n_i} \boldsymbol{F}_{iv}(\theta) \Big\} \mathbf{1}, \quad k \neq l \ \text{ and } \ m \in \{1,\ldots,s\}, \\
\frac{\partial}{\partial \phi} L_i(\theta) &= \sum_{j=1}^{n_i} \pi_i' \Big\{ \prod_{v=1}^{j-1} \boldsymbol{F}_{iv}(\theta) \Big\} \boldsymbol{D}_{\phi}^{(ij)}(\theta) \Big\{ \prod_{v=j+1}^{n_i} \boldsymbol{F}_{iv}(\theta) \Big\} \mathbf{1}, \\
\boldsymbol{D}_{\beta_{mk}}^{(ij)}(\theta) &= \Big\{ \frac{\partial}{\partial \beta_{mk}} \log\big(f(Y_{ij}; X_{ij}, \beta_k, \phi)\big) \Big\} \boldsymbol{F}_{ij}(\theta) e_k e_k', \\
\boldsymbol{D}_{\phi}^{(ij)}(\theta) &= \boldsymbol{F}_{ij}(\theta) \Big\{ \frac{\partial}{\partial \phi} \ln\big(\tilde{\boldsymbol{Y}}_{ij}(\theta)\big) \Big\}, \\
\boldsymbol{D}_{\gamma_{klm}}^{(ij)}(\theta) &= Z_{im} g_{ikl} \tilde{t}_{ij} \int_0^1 \exp\{\boldsymbol{G}_i \tilde{t}_{ij} u\}(e_k e_l' - e_l e_k') \exp\{\boldsymbol{G}_i \tilde{t}_{ij}(1-u)\} du \exp\{-\boldsymbol{G}_i \tilde{t}_{ij}\} \boldsymbol{F}_{ij}(\theta),
\end{aligned} \tag{7}$$

where $\prod_a^b$ is equal to the identity matrix if $a > b$, and the integral of the exponential matrix in $\boldsymbol{D}_{\gamma_{klm}}^{(ij)}(\theta)$ can be efficiently calculated by (6).

The main difficulty of the maximization is the evaluation of the exponential matrices. There are several ways to calculate the exponential of a matrix [29]. For simplicity, we can use $(I+\boldsymbol{A}/N_0)^{N_0}$ with a sufficiently large N_0 to approximate $\exp(\boldsymbol{A})$ [31]. Notice that each $L_i(\theta)$ is a product of n_i number of such exponential matrices, the resulting approximation error may become inflated. We approximate $\exp(\boldsymbol{G}_i\tilde{t}_{ij})$ using $(I + \boldsymbol{G}_i\tilde{t}_{ij}/N_i)^{N_i}$ with $N_i = n_i N_0$ and set $N_0 = 10\,000$ to obtain a good approximation of $L_i(\theta)$ in this study. The ML estimator of θ can be obtained by using the optimization function *fminunc* of MATLAB with provided log-likelihood function $-L$, gradient vector $-\dot{L}$, and Hessian matrix $-H$. Pseudocode for calculating L, $\dot{L}$, and H is shown as follows:

Algorithm Calculating the Likelihood, Gradient Vector and Hessian matrix

```
Input: θ, observed data
for i = 1 to n do
    L_i = I, N_i = 10^4 n_i.
    for j = 1 to n_i do
        Approximate exp(G_i t̃_ij) by P̂_ij = (I + G_i t̃_ij/N_i)^{N_i},   L_i* = P̂_ij * Ỹ_ij
    end for
    L_i = π_i' L_i 1,  using the formulas in (7) to calculate L̇_i = ∂L_i/∂θ,
    L+ = log(L_i), L̇+ = L_i^{-1} L̇_i, H− = L_i^{-2} L̇_i^{⊗2}
end for
Output: L, L̇, H
```

The proposed estimation procedure directly maximizes $L(\theta)$ and avoids the use of the expectation–maximization (EM) algorithm for sampling hidden states through the forward and backward algorithm and time-consuming iterations. Thus, its implementation is simpler and computationally more efficient than the EM algorithm. We compare the proposed method and the EM algorithm in terms of estimation accuracy and computational efficiency in the setting of Simulation 1 with a server *Intel(R) Xeon(R) CPU (X5690 3.47* GHz*)*. The two methods produce almost identical estimation. However, the proposed method converges much faster than the EM algorithm. On the basis of 500 replications, the proposed and EM methods take on the average 1.9 and 7.0 s as well as 11 and 38 iterations, respectively, to converge when $Y_i(t)$ is normal and $n = 500$. The comparison results in other settings are similar and omitted.

3.2. Theoretical properties of ML estimators

Let θ_0 be the true value of θ and $\hat{\theta}$ be the obtained ML estimator of θ. The asymptotic normality of $\hat{\theta}$ is not obvious because of the complexity of $L(\theta)$. We present the result in the following theorem and provide the proof in the Appendix.

Theorem 1. *Under the regularity conditions (C1)–(C4) provided in the Appendix, for $d = d_0$, $\sqrt{n}(\hat{\theta}-\theta_0)$ converges in distribution to a normal variable with mean 0 and variance Σ^{-1}, which can be consistently estimated by $\hat{\Sigma}^{-1}$, where*

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} L_i(\hat{\theta})^{-2}\dot{L}_i(\hat{\theta})\dot{L}_i'(\hat{\theta}), \quad \dot{L}_i(\theta) = \partial L_i(\theta)/\partial\theta.$$

Notably, this theorem is derived for the traditional longitudinal data, in which the number of observations per unit is bounded in probability. For the situation that $\max_{1\le i\le n} n_i \to \infty$, the convergence rate could be improved and further investigation is required.

With the estimated parameters, we can recover the trajectory of the hidden Markov process for each observed subject through the "Maximum a Posteriori" path. Given the observation $\mathcal{O}_i = \{W_i, Z_i, t_{ij}, X_{ij}, Y_{ij}, j \in \{1, \ldots, n_i\}\}$ of subject i, the distribution of the initial state is $\hat{\pi}_i$ and the transition probability matrix from $t_{i,j-1}$ to t_{ij} is $\hat{P}^{(i)}(\tilde{t}_{ij})$, where $\hat{P}^{(i)}(s) = \exp\{\hat{\boldsymbol{G}}_i s\}$. The posterior distribution of the state at observation time $\delta_{ij}(k) = P(S_i(t_{ij}) = k|\mathcal{O}_i)$ can be calculated using the forward–backward procedure [25,30]. Therefore, we can estimate the state $S_i(t_{ij})$ at time t_{ij} by using $\hat{S}_{ij} = \text{Argmax}_{1\le k\le d}\ \delta_{ij}(k)$. For the initial state, we take $\hat{S}_{i0} = \hat{S}_{i1}$ if $t_{i1} = 0$ and $\hat{S}_{i0} = \text{Argmax}_{1\le k\le d}\ \hat{\pi}_i(k)$ otherwise. For a general time t, if $t_{ij_0} < t < t_{i,j_0+1}$ for some $0 \le j_0 < n_i$, we can estimate $S_i(t)$ as follows:

$$\hat{S}_i(t) = \underset{1\le k\le d}{\text{Argmax}}\hat{P}^{(i)}_{uk}(t - t_{ij_0})\hat{P}^{(i)}_{kv}(t_{i,j_0+1} - t),$$

where $u = \hat{S}_{ij_0}$, $v = \hat{S}_{i,j_0+1}$, and $\hat{P}^{(i)}_{lk}(s)$ is the (l,k)th element of the matrix $\hat{P}^{(i)}(s)$. For subject i at future time $t > t_{in_i}$, we can predict the hidden state $S_i(t)$ by using $\hat{S}_i(t) = \text{Argmax}_{1\le k\le d}\hat{P}^{(i)}_{\hat{S}_{in_i},k}(t - t_{in_i})$.

For the assessment of model fitting, we define a pseudo-residual as follows:

$$\hat{r}_i = n_i^{-1}\sum_{j=1}^{n_i}\hat{V}_{ij}^{-1/2}\left\{Y_{ij} - \sum_{k=1}^{d}\delta_{ij}(k)\hat{\mu}_k(X_{ij})\right\},$$

where $\hat{\mu}_k(X_{ij})$ is estimation of $E(Y_{ij}|X_{ij}, S_i(t_{ij}) = k)$ and $\hat{V}_{ij}$ is the posterior variance of Y_{ij}. If the model fits the data adequately, $\hat{r}_i$s should be randomly fluctuated around zero. Hence, a scatter plot of residuals can provide information about the goodness-of-fit of the posited model.

Remark 2. The preceding ML procedure assumes that the proposed model is fully identifiable. However, label switching may occur in HMMs because the likelihood function is invariant to a permutation of state labels. Let d_0 be the true number of hidden states of an HMM, which is actually equal to the number of distinct values of $\beta_1, \ldots, \beta_d$. If $d = d_0$, the above label switching is not a problem because it affects only the label ranking. If $d > d_0$, then the HMMs and the associated parameters may become unidentifiable. We discuss this identifiability issue and the proposed solution in Section 4.

4. Number of hidden states

In practice, d_0 is usually unknown and need to be estimated. if $d > d_0$, then the state-specific parameter vectors $\beta_1, \ldots, \beta_d$ do not need to be distinct in the estimation procedure. Consequently, two or more states share the same parameters [27]. This section considers two relevant issues. Section 4.1 discusses the identification problem of model parameters in the case of $d > d_0$. Section 4.2 proposes a new penalized procedure for the selection of d_0 when knowing that $d_0 < d$ for some given d.

4.1. Identification of model parameters

A central question is how to identify model parameters for an HMM when $d > d_0$? If two states, for example, state k and state l, have the same parameter $\beta_k = \beta_l$, they are essentially the same state. Therefore, they should share the same characteristics such as between-transition probabilities with other states. However, the initial probabilities of states k and l can be arbitrary with a fixed summation. In the following theorem, we provide a sufficient condition to guarantee the equality of the two states.

Theorem 2. *For a continuous time HMM with d states and parameter θ, splitting state k_0 into two states k_1 and k_2 results in a new HMM with $d+1$ states and parameter $\theta^{(new)}$. The likelihood function (4) evaluated at θ and $\theta^{(new)}$ will be equal to each other if one takes*

(i) $\beta_{k_1}^{(new)} = \beta_{k_2}^{(new)} = \beta_{k_0}$.
(ii) $\exp\{\eta_{k_1}^{(new)'} W\} + \exp\{\eta_{k_2}^{(new)'} W\} = \exp\{\eta_{k_0}' W\}$ *for any given* W, *where* ηs *and* W *are similarly defined as those in* (2).
(iii) $g_{ik_1k}^{(new)} = g_{ik_2k}^{(new)} = g_{ik_0k}$ *and* $g_{ikk_1}^{(new)} + g_{ikk_2}^{(new)} = g_{ikk_0}$, *for* $k \notin \{k_0, k_1, k_2\}$, *where* $g_{ikl}^{(new)} = \exp\{Z_i' \gamma_{kl}^{(new)}\}$, *for* $k \neq l$.

In Theorem 2, Condition (i) ensures that the new states k_1 and k_2 share the same state-specific parameters. Condition (ii) implies that the initial probabilities of the new states k_1 and k_2 satisfy $\pi_{ik_1} + \pi_{ik_2} = \pi_{ik_0}$. Since the instantaneous transition probability is $P_{kl}(h) = hg_{kl} + o(h)$ when $h \to 0$ for $k \neq l$, Condition (iii) implies that instantaneous transition probabilities between other states and the new states k_1 and k_2 are the same as those between other states and state k_0. The proof is given in the Appendix.

4.2. Selection of the number of hidden states

We propose a new penalized likelihood method to select the number of hidden states of an HMM, d_0. Let d denote the upper bound of the number of hidden states. Naturally, a penalization should be introduced to make the same states to have parameters satisfying conditions in Theorem 2. However, implementing such a penalization is difficult because of label switching which prevents us from fixating the states. To avoid the label switching problem, we rank the state-specific parameters $\beta_1, \ldots, \beta_d$ such that all the states are well labeled and the same states are adjoining. Assume that β_1 has the minimum Euclid norm and β_2 has the smallest Euclid distance to β_1 among $\beta_2, \ldots, \beta_d$. In general, β_{k+1} has the smallest Euclid distance to β_k among $\beta_k, \ldots, \beta_d$. That is, $\|\beta_{k+1} - \beta_k\| \leq \|\beta_l - \beta_k\|$ for $l > k+1$ and $k \in \{1, \ldots, d-1\}$, where $\|\cdot\|$ denotes the Euclid norm. Notably, two different vectors may have the same Euclid norm. Thus, in the preceding norm ranking procedure, if two vectors have the same norm, then we compare them coordinatewisely. For example, if $\|\beta_1\| = \|\beta_2\|$, we label the state of β_1 as state 1 if there exists a $k_0 \leq d$ such that $\beta_{k1} \leq \beta_{k2}$ for all $k \leq k_0$. This ordering implies that $d > d_0$ is equivalent to $\|\beta_{k+1} - \beta_k\| = 0$ for some k. We define $\sigma_k^2 = \left\{\|\beta_{k+1} - \beta_k\|^2 + \|\eta_{k+1} - \eta_k\|^2 + \|\gamma_{k+1,k} - \gamma_{k,k+1}\|^2 + \sum_{l \neq k+1,k}\{\|\gamma_{k+1,l} - \gamma_{kl}\|^2 + \|\gamma_{l,k+1} - \gamma_{lk}\|^2\}\right\}/d_{pen}$, where $d_{pen} = p+q+s+2s(d-2) = p+q+s(2d-3)$ is the total number of summations. We note from (2) that $\pi_{ik} = \exp\{\eta_k' W_i\}\pi_{id}$. Thus, $\|\eta_{k+1} - \eta_k\| = 0$ implies that $\pi_{i,k+1} = \pi_{i,k}$. Consequently, a penalization on σ_k enforces $\beta_k = \beta_{k+1}$ and $\pi_{ik} = \pi_{i,k+1}$ if k and $k+1$ are the same state. Based on Theorem 2, $\sigma_k = 0$ implies that states k and $k+1$ are the same.

A penalized log-likelihood function is defined as

$$\tilde{L}(\theta) = L(\theta) - N \sum_{k=1}^{d-1} p_{\lambda_N}(\sigma_k), \tag{8}$$

where $N = \sum_{i=1}^n n_i$ is the total number of observations, and λ_N is a tuning parameter. We select the penalty function p_λ to be the smoothly clipped absolute deviation penalty (SCAD) [11] with its derivative defined as $\dot{p}_\lambda(x) = \lambda I(|x| \leq \lambda) + \frac{(a\lambda - |x|)_+}{a-1} I(|x| > \lambda)$. This penalization method provides a fused learning of the multivariate state parameters via proper ordering. Chen and Khalili [10] and Hung et al. [17] proposed a double penalized log-likelihood function in selection of the number of hidden states of finite mixture models and HMMs as $\tilde{L}_{DP}(\theta) = L(\theta) + C_d \sum_{k=1}^d \log(\pi_k) - N \sum_{k=1}^{d-1} p_{\lambda_N}(\beta_{k+1} - \beta_k)$, where π_k and β_k, $k \in \{1, \ldots, d\}$ are similarly defined as π_{ik} and β_k in the proposed model, and C_d and λ_N are tuning parameters to separately penalize π_k and $\beta_{k+1} - \beta_k$. Despite its successful applications in the aforementioned literature, $\tilde{L}_{DP}(\theta)$ assumed homogeneous π_k and one-dimensional β_k, and disregarded the dynamic transition feature of HMMs. In the proposed continuous-time HMMs, the probabilities $\pi_{ik}, i \in \{1, \ldots, n\}, k \in \{1, \ldots, d\}$ and the elements of the generating matrix $g_{ikj}, i \in \{1, \ldots, n\}, k \in \{1, \ldots, d\}, j \in \{1, \ldots, d\}$ depend on covariate vectors W_i and Z_i, respectively, and between-state transitions are associated with a multi-dimensional parameter vector of β_k, η_k, and γ_{kl}. Thus, the above double penalized procedure, which was originally designed for mixture models without dynamic state transition, is not directly applicable to the current continuous-time HMMs. By contrast, the proposed approach simultaneously penalizes the differences of adjacent β_k, η_k, and γ_{kl}, $k, l \in \{1, \ldots, d\}$, thereby accommodating the dynamic transition feature. Using the idea of the double penalized procedure, we may allow distinct tuning parameters to separately penalize the differences of adjacent β_k, γ_{kl}, and η_k. However, the involved computational burden will increase dramatically because selection of the three tuning parameters is extremely time-consuming. In comparison, the proposed procedure includes only one tuning parameter and is thus computationally more efficient.

Let $\bar{\theta}$ be the maximizer of the penalized log-likelihood function $\tilde{L}(\theta)$ and denote the corresponding estimators of state-specific parameters as $\bar{\beta}_k$. The estimator of the number of hidden states of HMM, $\hat{d}_0$, is defined as

$$\hat{d}_0 = \text{the number of distinct values of } \{\bar{\beta}_k,\ k \in \{1, \ldots, d\}\}.$$

The following theorem provides the consistency of the resulting estimator $\hat{d}_0$ with the proof provided in the Appendix.

Theorem 3. *Under the regular conditions of Theorem 1, if* $\lambda_N \to 0$ *and* $\sqrt{n}\lambda_N \to \infty$, *then the penalized estimator of the number of hidden states* $\hat{d}_0$ *for* $d_0 < d$ *converges in probability to the true number of hidden states* d_0 *of HMM.*

Table 1
Estimation results based on 1000 replications for the continuous time HMM in Simulation 1, where β_{jk} and η_k are the regression coefficients in state $k, k \in \{1, 2\}$. Normal and Poisson distributions are considered for the responses. Bias is the sampling mean of the estimate minus the true value, SE is the sampling standard error of the estimator, SEE is the sampling mean of standard error estimate, and CP is the 95% empirical coverage probability.

Para.	$n = 100$				$n = 200$				$n = 500$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
Normal Distribution												
β_{11}	0.000	0.065	0.073	0.953	−0.001	0.046	0.049	0.951	0.001	0.030	0.030	0.939
β_{21}	−0.001	0.094	0.108	0.961	0.002	0.071	0.072	0.942	−0.001	0.042	0.043	0.948
β_{31}	0.000	0.016	0.018	0.967	−0.000	0.012	0.012	0.943	−0.000	0.007	0.007	0.945
β_{12}	0.001	0.073	0.081	0.951	0.002	0.053	0.054	0.931	−0.001	0.033	0.033	0.940
β_{22}	0.000	0.101	0.117	0.964	−0.000	0.073	0.076	0.955	0.002	0.046	0.046	0.934
β_{32}	0.000	0.019	0.022	0.962	−0.000	0.013	0.014	0.955	0.000	0.008	0.009	0.956
η_1	0.007	0.376	0.390	0.973	0.008	0.259	0.263	0.957	0.005	0.164	0.161	0.946
γ_{12}	−0.045	0.328	0.346	0.976	−0.002	0.225	0.228	0.946	−0.005	0.140	0.139	0.942
γ_{21}	0.003	0.240	0.271	0.965	−0.000	0.180	0.182	0.951	−0.006	0.113	0.112	0.945
σ^2	−0.004	0.014	0.015	0.943	−0.002	0.010	0.010	0.947	−0.001	0.006	0.006	0.954
Poisson Distribution												
β_{11}	−0.006	0.109	0.125	0.970	−0.002	0.076	0.084	0.962	0.003	0.050	0.051	0.953
β_{21}	−0.003	0.165	0.181	0.964	−0.004	0.111	0.120	0.957	0.002	0.069	0.072	0.948
β_{31}	−0.001	0.027	0.030	0.967	0.000	0.018	0.019	0.948	−0.001	0.011	0.012	0.965
β_{12}	0.001	0.075	0.088	0.951	−0.001	0.054	0.059	0.951	−0.001	0.034	0.036	0.954
β_{22}	−0.000	0.094	0.117	0.974	−0.001	0.068	0.075	0.959	−0.000	0.043	0.045	0.951
β_{32}	−0.001	0.017	0.020	0.959	0.000	0.012	0.013	0.953	0.000	0.008	0.008	0.961
η_1	0.021	0.416	0.424	0.975	0.002	0.271	0.285	0.958	0.009	0.176	0.176	0.955
γ_{12}	−0.039	0.347	0.370	0.970	−0.013	0.251	0.245	0.958	−0.008	0.151	0.150	0.950
γ_{21}	−0.030	0.275	0.290	0.967	−0.018	0.189	0.194	0.951	−0.007	0.118	0.118	0.940

Remark 3. As suggested by [11], we use a local quadratic approximation for the penalty function. That is, for x close to x_0, we take $p_\lambda(x) = p_\lambda(x_0) + \frac{\dot{p}_\lambda(x_0)}{2x_0}(x^2 - x_0^2)$. For the tuning parameters, we take $a = 3.7$ and select λ_N based on BIC. Specifically, let $\bar{\theta}_\lambda$ and $\hat{d}_\lambda$ be the estimators that correspond to λ, then we select λ_N by minimizing $BIC(\lambda) = -2L(\bar{\theta}_\lambda) + \log(N)D_\lambda$, where $D_\lambda = \hat{d}_\lambda(s(\hat{d}_\lambda - 1) + p + q) - q + r$ is the number of effective parameters given $\hat{d}_\lambda$ states. When d is large, the alternating direction method of multipliers [7] can be used for the optimization.

Remark 4. In practice, a rigorous ordering of the states may be impossible because the true values of $\beta_1, \ldots, \beta_d$ are unknown. The following steps are suggested to initialize the estimation procedure: (i) We obtain the ML estimators $\beta^{(0)}$ and $\phi^{(0)}$ by assuming only one hidden state. (ii) We set $\beta_k^* = \beta^{(0)}(1 + c\epsilon_k)$, $k \in \{1, \ldots, d\}$, where ϵ_k are the disturbances that are independent and distributed as $N(0, \boldsymbol{I}_p)$, and c is a small number with a commonly suggested value of $c = 0.1$. (iii) We assign the initial values $\beta_{(k)}^{(0)}$, $k \in \{1, \ldots, d\}$ to be the ordered array of $\{\beta_k^*, k \in \{1, \ldots, d\}\}$, $\gamma_{kl}^{(0)} = 0$, $k \neq l$, and $\eta_k^{(0)} = 0, k \in \{1, \ldots, d-1\}$. This initialization method works well in our simulation studies.

5. Simulation study

5.1. Simulation 1

We first considered a continuous-time HMM with $d_0 = 2$, in which the response variable $Y_i(t)$ in state k was generated from (1) a normal distribution with mean $\beta_k' X_i(t)$ and variance $\sigma^2 = 0.25$ and (2) a Poisson distribution with mean $\exp\{\beta_k' X_i(t)\}$. We set $X_i(t) = (X_{i1}, X_{i2}, t)$, where X_{i1} and X_{i2} were generated from a Bernoulli distribution with success probability 0.5 and a uniform distribution on (0, 1), respectively. In both cases, the state-specific covariate effects were set as $\beta_1 = (\beta_{11}, \beta_{21}, \beta_{31})' = (-1, 0.5, 0.2)'$ and $\beta_2 = (\beta_{12}, \beta_{22}, \beta_{32})' = (1, 0.5, 0.2)'$. The observation time t was generated from a Poisson process with intensity $\exp\{0.05 + 0.5X_{i1}\}$ and was censored by time $C_i = \min(C_i^*, \tau)$, where $\tau = 6$ and C_i^* follows a uniform distribution on interval [3, 8]. The average observations of each subject is about 7. For the hidden Markov process, we take $Z_i \equiv 1$ and set $\gamma_{12} = -2$ and $\gamma_{21} = -1.5$. For the initial state probability, we set $W_i = 1$ and $\eta_1 = -0.3$. Sample sizes $n = 100$, 200, and 500 were considered for each case.

The estimation results based on 1000 replications are reported in Table 1, where Bias is the sampling mean of the estimate minus the true value, SE is the sampling standard error of the estimator, SEE is the sampling mean of standard error estimate, and CP is the 95% empirical coverage probability. The biases of the parameter estimates are small, their estimated standard errors are close to the sampling standard errors, and the CP's are close to the nominal level. As expect, the parameter estimates perform better for the regression model than for the transition model, and the overall performance is improved when the sample size increases.

Table 2
Selection proportions of the number of hidden states based on 400 replications in Simulation 2, where $d_0 \in \{2, 3\}$ is the true number of states and the dimension of X is $p = 3$. Normal and Poisson distributions are considered for the responses. 'Pen', 'AIC' and 'BIC' stand for the proposed penalization method, the AIC criterion and the BIC criterion, respectively.

Criterion	d_0	$\hat{d}$	Normal			Poisson		
			$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
Pen	2	2	0.9925	0.9975	1.0000	0.9825	0.9850	0.9600
		3	0.0075	0.0025	0	0.0175	0.0150	0.0400
	3	2	0.1950	0.0175	0	0.3475	0.0650	0
		3	0.8050	0.9800	0.9950	0.6525	0.9225	0.9450
		4	0	0.0025	0.0050	0	0.0125	0.0550
AIC	2	2	0.9825	0.9825	0.9800	0.9775	0.9825	0.9825
		3	0.0175	0.0175	0.0200	0.0200	0.0175	0.0175
		4	0	0	0	0.0025	0	0
	3	2	0.0775	0.0125	0.0050	0.2125	0.0250	0.0075
		3	0.8050	0.8875	0.9100	0.6925	0.8175	0.8950
		4	0.1175	0.1000	0.0850	0.0950	0.1575	0.0975
BIC	2	2	1.0000	1.0000	1.0000	0.9950	1.0000	1.0000
		3	0	0	0	0.0050	0	0
	3	2	0.5750	0.0925	0.0100	0.6875	0.2075	0.0325
		3	0.4250	0.8950	0.9350	0.3125	0.7825	0.8975
		4	0	0.0125	0.0550	0	0.0100	0.0700

5.2. Simulation 2

To evaluate the proposed penalized procedure in selecting the number of hidden states of HMMs, we considered two cases of d_0 = 2 and 3. The model parameters were set as follows: for $d_0 = 2$, $\beta_1 = (-1, 0.5, 0.2)'$, $\beta_2 = (1, 0.5, 0.2)'$, $\eta_1 = -0.3$, $\gamma_{12} = -2$, and $\gamma_{21} = -1.5$; for $d_0 = 3$, $\beta_1 = (0, 0.5, 0.2)'$, $\beta_2 = (1, 1, 0.2)'$, $\beta_3 = (2, 0.5, 0.2)'$, $\eta_1 = 0.2$, $\eta_2 = -0.3$, $\gamma_{12} = -0.5$, $\gamma_{13} = -2$, $\gamma_{21} = -1.5$, $\gamma_{23} = 0$, $\gamma_{31} = -1.5$, and $\gamma_{32} = -1$. The turning parameter λ_N was selected according to Remark 3 by searching $\lambda = cN^{-1/2}\log(N)$ within interval $c \in [0.05, 1]$. For comparison, we also determined the order using AIC and BIC. The results based on 400 replications are reported in Table 2, in which each value is a proportion of the selected number of hidden states. When $d_0 = 2$, the proposed method performs similarly to AIC and BIC. When $d_0 = 3$, the proposed method outperforms AIC and BIC, especially for a relatively large n, say $n = 200$ or 500. The proportions of selecting correct number of hidden states by the proposed method are close to 1 when n increases, thereby reconfirming the consistency of the order estimator shown in Theorem 3.

To further examine (a) whether the proposed approach can handle larger number of hidden states and/or more baseline covariates and (b) how different transition structures affect the estimation performance, we considered a larger model with $d_0 = 5$, $p = 6$, and two types of generating matrices as follows: (1) a general generating matrix $\boldsymbol{G}_1$ and (2) a band generating matrix $\boldsymbol{G}_2$ that only allows transitions between certain adjacent states. The model setup and simulation results are provided in Tables S1–S3 of the supplementary material. As expected, the parameter estimates, especially those involved in the generating matrix, perform better under $\boldsymbol{G}_2$ than under $\boldsymbol{G}_1$. The proposed method consistently outperforms AIC and BIC, and its performance improves when the sample size increases.

5.3. Simulation 3

We evaluated the performance of the MAP method in recovering the trajectory of the hidden Markov process on the basis of the previous simulation. We considered the setup of Simulation 1 except for resetting $\beta_1 = (-1, -0.5, -0.2)$ to distinguish the two states when $X_{i1} = 0$. Considering that each subject may have different elapsed times, we evaluated the performance of the MAP method by using the proportion of times when the state of a subject is correctly estimated. Let $I_{ij} = I(\hat{S}_i(\kappa_j t_{in_i}) = S_i(\kappa_j t_{in_i}))$, $\kappa_j = j/100, j \in \{0, 1, 2, \ldots, 120\}$, and define $CR_{ki} = 33^{-1}\sum_{j:(k-1)/3<\kappa_j\le k/3} I_{ij}$ for $k \in \{1, 2, 3\}$, $CR_{Ti} = 101^{-1}\sum_{j=0}^{100} I_{ij}$, and $CR_{Pi} = 20^{-1}\sum_{j=101}^{120} I_{ij}$. Hence, CR_{1i}, CR_{2i}, CR_{3i}, and CR_{Ti} represent the proportions of correctly estimated hidden states of subject i on the first, second, and the last one-third of the whole elapsed time interval, respectively. CR_{Pi} evaluates the prediction accuracy within the extended 20% future time interval of subject i. Table 3 reports all the proportions calculated based on 500 replications. The proposed method works satisfactorily in recovering and predicting the trajectory of hidden Markov process. The performance improves over time and the number of observations per unit.

6. Application to bladder cancer data

In this section, we applied the proposed method to the longitudinal bladder tumor study conducted by the Veterans Administration Cooperative Urological Research Group [35]. In the study, 85 patients with superficial bladder tumors

Table 3
Mean proportions of correctly estimated hidden states based on 500 replications in Simulation 3, where n_i is the number of observation of the ith unit. CR_1, CR_2, CR_3 and CR_T represent the mean proportions of correctly estimated hidden states of subject i on the first, second, last one-third and the total of the whole elapsed time interval, respectively. CR_P evaluates the mean prediction accuracy within the extended 20% future time interval of each subject.

	Size of n_i	CR_1	CR_2	CR_3	CR_T	CR_P
Normal	$n_i \geq 3$	0.897	0.922	0.949	0.923	0.921
	$n_i \geq 6$	0.917	0.939	0.957	0.938	0.921
	$n_i \geq 9$	0.943	0.955	0.965	0.954	0.920
	$n_i \geq 12$	0.962	0.965	0.970	0.965	0.915
Poisson	$n_i \geq 3$	0.842	0.881	0.918	0.880	0.891
	$n_i \geq 6$	0.869	0.906	0.937	0.904	0.902
	$n_i \geq 9$	0.908	0.936	0.955	0.933	0.911
	$n_i \geq 12$	0.940	0.955	0.966	0.953	0.913

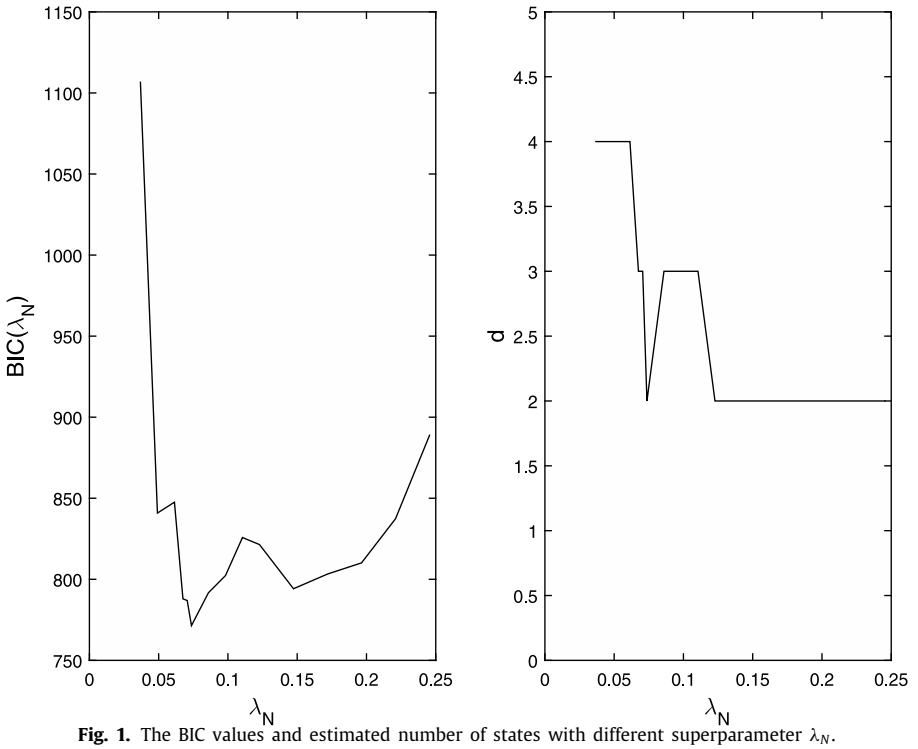


Fig. 1. The BIC values and estimated number of states with different superparameter λ_N.

were randomly assigned to the thiotepa treatment group and the placebo group. All tumors (if any) were transurethrally removed during each clinical visit, and many patients experienced multiple tumor recurrences during the study. The number of clinical visits and the number of bladder tumors occurrence between clinical visits for each patient were recorded. Apparently, the number of tumors exhibited heterogeneity (see Fig. S1 of the supplementary material). The total number of clinical visits for each patient varies from 2 to 39 with a mean of 12, and the longest observation time is 4.42 years. Two baseline covariates, namely, the number of initial tumors before entering the study and the size of the largest initial tumor, were measured. This study investigates how the number of tumors changes over time and its dynamic heterogeneity.

Let $Y_i(t)$ be the number of observed tumors at time t and $Y_i(0)$ be the initial number of tumors. We assumed that $Y_i(t)$ followed a Poisson distribution with mean $\exp\{X_i(t)'\beta\}$. We first considered a simple case of $d = 1$ with $X_i(t) = (X_{i1}, X_{i2}, t, t^{1/2}, t^2, t^3)'$, where $X_{i1} = 1$ and X_{i2} is the treatment indicator, to obtain a rough idea of the parametric form of t in $X_i(t)$. The coefficient estimates, together with their standard error estimates in parentheses, were 0.996(0.066), -0.718(0.034), 2.652(0.677), -4.141(0.500), -0.149(0.240), and -0.029(0.034). These estimates suggested that the quadratic and cubic terms t^2 and t^3 were unnecessary. Thus, we took $X_i(t) = (1, X_{i2}, t, t^{1/2})'$. We also tried several choices of W_i and Z_i and disregard the components with nonsignificant coefficients. Finally, we set W_i as the size of the largest initial tumor to model the initial states and Z_i as the treatment indicator to model the generating matrix.

Fig. 1 provides numerical evidences of the model selection procedure. The left panel presents the plots of BIC values corresponding to the turning parameter λ_N, whereas the right panel depicts the number of hidden states d versus λ_N.

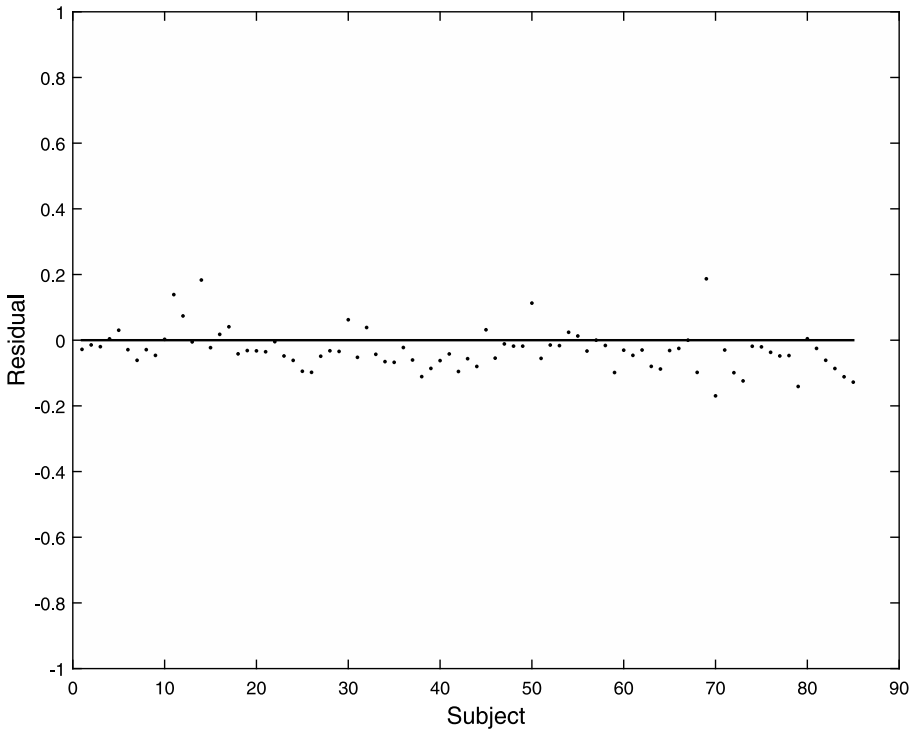


Fig. 2. Scatter plot of the residuals of the bladder tumor data.

Apparently, a two-state HMM was selected. Fig. 2 presents the scatter plot of the pseudo-residuals defined in Section 3.2. The pseudo-residuals distribute randomly around zero, thereby suggesting that the selected model fits the data well. Table 4 reports the parameter estimates of the selected model. The estimated treatment effect is in agreement with those reported in the existing literature [3,24,35–37]. However, our study provides new insights into the pathophysiological states of bladder cancer, the progression of the disease, and the state-specific effect of thiotepa treatment. The intercept is $\hat{\beta}_{11} = 1.9546(0.2157)$ in state 1 and $\hat{\beta}_{12} = 0.6007(0.1396)$ in state 2. Thus, states 1 and 2 can be regarded as deterioration and amelioration states, respectively. On average, patients had substantially more tumors in the deterioration state than in the amelioration state. Fig. S2 plots the estimated mean number of tumors versus time for the treatment group. Apparently, the number of tumors decreased at the early stage of treatment and then increased over time in state 1 but remained stable in state 2. In both states, the thiotepa treatment significantly reduced the number of tumors, and this effect became more pronounced in the deterioration state than in the amelioration state. Moreover, recalling that $g_{ikk} = -\sum_{k\neq l} g_{ikl} = -\sum_{k\neq l} \exp\{\hat{\gamma}_{kl} Z_i\}$ and that the duration time of adjacent occurrences of tumors in state k follows the exponential distribution with mean $-g_{ikk}^{-1}$, we calculated from Table 4 that patients in the treatment group ($Z_i = 1$) would stay on average 0.35 and 3.01 years in states 1 and 2, respectively, whereas patients in the placebo group ($Z_i = 0$) would stay on average 1 year in both states. For a patient in state 1 of the treatment group, after approximately 4 months of treatment, the disease might become ameliorated and lasted around 3.01 years in state 2. We also calculated the transition probability matrices $\boldsymbol{P}(t) = \exp(\boldsymbol{G}t)$ for $t = j/12$, where $j \in \{1, 3, \ldots, 23\}$ ranges from one month to two years. The results are depicted in Fig. S3 and reported in Table S4. We observed that after approximately one year, the transition probability matrix $\boldsymbol{P}(t)$ converged with stable distributions (0.11, 0.89) and (0.5, 0.5) for the treatment and placebo group, respectively. That is, after about one year, approximately 89% patients in the treatment group, while about 50% patients in the placebo group remained in state 2 (amelioration). These results confirm the effectiveness of the treatment. Fig. 3 presents an intuitive overview of these results with plots of the observed numbers of tumors and estimated individual hidden trajectories of four randomly selected patients.

Notably, state 2 represents an amelioration state, in which the mean number of tumors is small. Baetschmann and Winkelmann [3] adopted a zero-inflated model to analyze this data and identified cured subjects in the recurrence of bladder tumors. Thus, state 2 of our model might be connected with $Y = 0$ of the zero-inflated model. However, the estimated trajectories in Fig. 3 indicate that state 2 not only includes $Y = 0$ but also contains Ys of small magnitude. HMMs with an absorbing state may be considered to accommodate the zero-inflation feature. More efforts are required for further investigation in the future.

Table 4
Estimation results for the bladder cancer data with 2 hidden states, where 'Par.' stands for the parameter, 'Est' stands for the estimate and 'Std' is the standard error estimate.

Estimate of β			Estimate of $\boldsymbol{\gamma}$		Estimate of η	
	State: 1	State: 2				
Covariates	Est(Std)	Est(Std)	Par.	Est(Std)	Par.	Est(Std)
1	1.9546 (0.2157)	0.6007 (0.1396)	γ_{12}	1.0358 (0.2764)	η_1	−1.2350 (0.2809)
Treat.	−0.2909 (0.0833)	−0.3118 (0.2349)	γ_{21}	−1.1030 (0.3572)		
t	0.6024 (0.1717)	3.3221 (0.4332)				
$t^{1/2}$	−1.4246(0.3994)	−7.9478 (0.7881)				

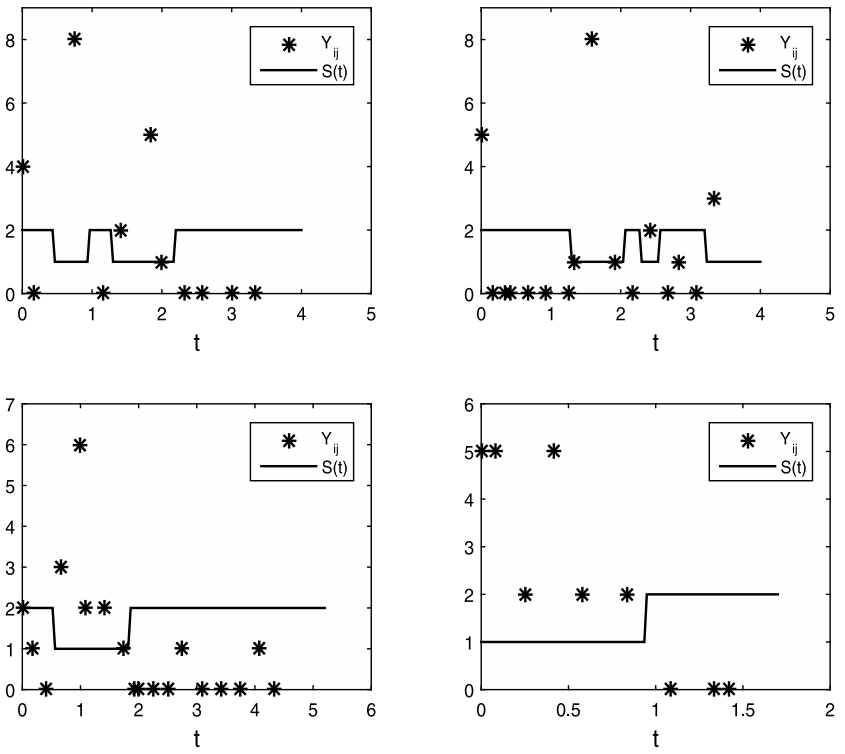


Fig. 3. The observed number of tumors and the estimated trajectories of 4 patients.

7. Conclusions

In this article, we proposed a continuous-time HMM to analyze longitudinal data. Unlike conventional discrete-time HMMs that restrict observation times to be discrete and regular, the proposed model allows observation time to be continuous and irregular, thereby enabling the dynamic heterogeneity of longitudinal data with non-predetermined observation times to be revealed. The proposed model subsumes discrete-time HMMs as a special case. Moreover, in contrast to conventional HMMs, the proposed model allows for a subject-specific generating matrix and regards the number of hidden states as an unknown parameter that must be estimated together with other unknowns. We developed an ML procedure coupled with an efficient algorithm to conduct estimation and a penalizing approach to select the number of hidden states. Finite sample behaviors of the proposed estimators of parameters and the number of hidden states are evaluated through simulation studies. The application of the proposed method to a study of a bladder tumor dataset provided new insights into the progression of bladder cancer.

This study has limitations. In the proposed model, we assumed that the duration of a visit to state k follows an exponential distribution with rate $-g_{kk}$. Given the unobservable nature of the transition process, this assumption is difficult to verify. Thus, a more general model without such an assumption is of considerable interest. For example, the inclusion of random effects can enhance the model capability of capturing more data features. The proposed model can be extended to include random effects at the cost of more demanding computation in parameter estimation. However, the theoretical results in selection of the number of hidden states may not hold when incorporating random effects into the proposed model. The feasibility of developing the theoretical results, such as Theorems 2 and 3, in the context of continuous-time mixed HMMs is uncertain and requires substantial efforts in the future. Moreover, we assumed a

homogeneous hidden Markov process. In practice, the generating matrix $\boldsymbol{G}$ may depend on time-dependent covariates or even on the history of the longitudinal variable. We may consider a transition model for $\boldsymbol{G}$ to deal with this feature. However, the proposed estimation procedure cannot be extended to accommodate non-homogeneous transitions in a straightforward manner, and the associated inference become extremely challenging. This interesting topic deserves further investigation. Note that the number of parameters increases fast with the number of hidden states and covariates, the proposed method may therefore not work well when the number of hidden states is large and/or covariates are of high dimension. Finally, the proposed model is developed in a parametric framework. A comprehensive approach is to model the longitudinal process and the hidden Markov process in a semi-parametric or nonparametric manner. This concern can be partially overcome through linear approximation of polynomials or splines. While it certainly enhances model flexibility in longitudinal data analysis, such an extension in the context of continuous-time HMMs is challenging, especially under the current maximum likelihood-based inference framework. Substantial effort is required to address these challenges.

Acknowledgments

We thank the Editor, Associate Editor and three referees. This research was partially supported by the National Natural Science Foundation of China Grants (Nos. 11671275, 11471223, 11771431, 11690015 and 11926341), Key Laboratory of RCSDS, CAS, PR China (No. 2008DP173182), GRF grants 14303017 and 14301918 from the Research Grants Council of the Hong Kong Special Administrative Region, and the Academy for Multidisciplinary Studies of Capital Normal University, PR China.

Appendix. Appendix: Proof of Theorems

We impose the following regularity conditions:

(C1) The parameter space for θ with d_0 states, denoted by $\Theta \equiv \Theta_\beta \times \Theta_\eta \times \Theta_\gamma \times \Theta_\phi$, is compact.

(C2) The first and second derivatives of the logarithmic of density function $f(y; x, \tilde{\beta}, \phi)$ satisfy the equations

$$E_{\tilde{\beta},\phi}\Big[\frac{\partial \log f(Y; x, \tilde{\beta}, \phi)}{\partial a}\Big] = 0, \quad \text{for } a \in \{\tilde{\beta}_h, h \in \{1, \ldots, p\},\ \phi_m, m \in \{1, \ldots, r\}\}$$

and for $a, b \in \{\tilde{\beta}_h, h \in \{1, \ldots, p\},\ \phi_m, m \in \{1, \ldots, r\}\}$,

$$E_{\tilde{\beta},\phi}\Big[\frac{\partial \log f(Y; x, \tilde{\beta}, \phi)}{\partial a}\frac{\partial \log f(Y; x, \tilde{\beta}, \phi)}{\partial b}\Big] = -E_{\tilde{\beta},\phi}\Big[\frac{\partial^2 \log f(Y; x, \tilde{\beta}, \phi)}{\partial a \partial b}\Big].$$

(C3) There exists an open subset ω of $\Theta_\beta \times \Theta_\phi$ that contains the true parameter point $\{(\beta_{k0}', \phi_0')', k \in \{1 \ldots, d_0\}\}$ such that for almost all y the density $f(y; x, \tilde{\beta}, \phi)$ admits all third derivatives $(\partial f(y; x, \tilde{\beta}, \phi))/(\partial a \partial b \partial c)$ with $a, b, c \in \{\tilde{\beta}_h, h \in \{1, \ldots, p\},\ \phi_m, m \in \{1, \ldots, r\}\}$ for all $(\tilde{\beta}', \phi')' \in \omega$. Furthermore, there exist functions M_{abc} with finite expectations such that

$$\left|\frac{\partial^3}{\partial a \partial b \partial c} \log f(y; x, \tilde{\beta}, \phi)\right| \le M_{abc}(y), \text{ for all } (\tilde{\beta}', \phi')' \in \omega.$$

(C4) Suppose that the true state parameters $\{\beta_{k0}, k \in \{1, \ldots, d_0\}\}$ are distinct, and the Fisher information matrix $I(\theta) = E[\dot{L}_i(\theta)\dot{L}_i(\theta)']$ is finite and positive definite at the true parameter point θ_0 where $\dot{L}_i(\theta)$ is the score function.

We first introduce some notations and lemmas before giving the proof. Let $\boldsymbol{B} = (B_{kl})_{d\times d}$ and $\boldsymbol{D} = (D_{kl})_{(d+1)\times(d+1)}$ be transition probability matrices for d and $d+1$ states, respectively. Define a class of relationship '$\boldsymbol{B} \prec \boldsymbol{D}$' as

$$\boldsymbol{B} \prec \boldsymbol{D}: \text{ if } \begin{cases} B_{kl} = D_{kl}, & k, l \in \{1, \ldots, d-1\}; \\ B_{dl} = D_{dl} = D_{d+1,l}, & l \in \{1, \ldots, d-1\}; \\ B_{kd} = D_{kd} + D_{k,d+1}, & k \in \{1, \ldots, d-1\}. \end{cases}$$

Lemma A.1. *If $\boldsymbol{B}_1 \prec \boldsymbol{D}_1$ and $\boldsymbol{B}_2 \prec \boldsymbol{D}_2$, then $\boldsymbol{B}_1\boldsymbol{B}_2 \prec \boldsymbol{D}_1\boldsymbol{D}_2$*

Proof. For $i \in \{1, 2\}$, since $\boldsymbol{B}_i \prec \boldsymbol{D}_i$, we can rewrite $\boldsymbol{B}_i, \boldsymbol{D}_i$ in form of block matrices as

$$\boldsymbol{B}_i = \begin{pmatrix} \boldsymbol{B}_0^{(i)} & b_i \\ a_i & c_i \end{pmatrix}, \qquad \boldsymbol{D}_i = \begin{pmatrix} \boldsymbol{B}_0^{(i)} & b_i^{(1)} & b_i^{(2)} \\ a_i & d_{11}^{(i)} & d_{12}^{(i)} \\ a_i & d_{21}^{(i)} & d_{22}^{(i)} \end{pmatrix},$$

where $\boldsymbol{B}_0^{(i)}$ is a $(d-1)\times(d-1)$ matrix, a_i is a $1\times(d-1)$ row vector, b_i and $b_i^{(j)}$, $j \in \{1, 2\}$ are $(d-1)\times 1$ column vectors satisfying $b_i^{(1)} + b_i^{(2)} = b_i$, and $c_i, d_{kl}^{(i)}$, $k, l \in \{1, 2\}$ are positive numbers. Note that $\boldsymbol{B}_i, \boldsymbol{D}_i, i \in \{1, 2\}$ are all transition

probability matrices which must sum to 1 for each row. Therefore, $d_{11}^{(i)} + d_{12}^{(i)} = d_{21}^{(i)} + d_{22}^{(i)} = c_i$, combining this with equations $b_i^{(1)} + b_i^{(2)} = b_i$, it is easy to verify that $\boldsymbol{B}_1\boldsymbol{B}_2 \prec \boldsymbol{D}_1\boldsymbol{D}_2$.

Lemma A.2. *Let $\boldsymbol{G}$ and $\boldsymbol{G}^{(new)}$ be generating matrices for d and $d+1$ states and satisfy condition (iii) of Theorem 2 with $k_0 = k_1 = d$ and $k_2 = d+1$. Define $\boldsymbol{P} = \exp(\boldsymbol{G})$ and $\boldsymbol{P}^{(new)} = \exp(\boldsymbol{G}^{(new)})$. Then, $\boldsymbol{P} \prec \boldsymbol{P}^{(new)}$.*

Proof. By the properties of exponential matrices, we have that, $\boldsymbol{P} = \lim_{n\to\infty}(\boldsymbol{P}_n)^n$ and $\boldsymbol{P}^{(new)} = \lim_{n\to\infty}(\boldsymbol{P}_n^{(new)})^n$, where $\boldsymbol{P}_n = \boldsymbol{I}_d + n^{-1}\boldsymbol{G}$ and $\boldsymbol{P}_n^{(new)} = \boldsymbol{I}_{d+1} + n^{-1}\boldsymbol{G}^{(new)}$. By the constrain (1) for generating matrices, $\boldsymbol{P}_n$ and $\boldsymbol{P}_n^{(new)}$ must be transition probability matrices for n large enough. In addition, since $\boldsymbol{G}$ and $\boldsymbol{G}^{(new)}$ satisfy condition (iii) of Theorem 2 with $k_0 = k_1 = d$ and $k_2 = d+1$, it is easy to verify that $\boldsymbol{P}_n \prec \boldsymbol{P}_n^{(new)}$. Therefore, by Lemma A.1, $\{\boldsymbol{P}_n\}^n \prec \{\boldsymbol{P}_n^{(new)}\}^n$. Some simple arguments lead to the conclusion that the limits also have this relationship, that is, $\boldsymbol{P} \prec \boldsymbol{P}^{(new)}$.

Lemma A.3. *For any given distribution π and transition probability matrix $\boldsymbol{P}_1, \boldsymbol{P}_2$, we have*

$$\pi'\boldsymbol{P}_1[\frac{\partial \exp(-\boldsymbol{G}t)}{\partial g_{kl}}]\boldsymbol{P}_2\mathbf{1} = 0, \quad \text{for all } k \neq l. \tag{9}$$

Proof. Recalling that $g_{kk} = -\sum_{l\neq k} g_{kl}$, we have $\partial\boldsymbol{G}/\partial g_{kl} = e_k(e_l - e_k)'$. By the derivative formula (5), we have

$$\pi'\boldsymbol{P}_1[\frac{\partial \exp(-\boldsymbol{G}t)}{\partial g_{kl}}]\boldsymbol{P}_2\mathbf{1} = t\pi'\boldsymbol{P}_1\int_0^1 \exp(\boldsymbol{G}tu)(e_ke_l' - e_ke_k')\exp(\boldsymbol{G}t(1-u))ds\boldsymbol{P}_2\mathbf{1} = t\int_0^1 \pi'\tilde{\boldsymbol{P}}_1(u)e_k(e_l - e_k)'\tilde{\boldsymbol{P}}_2(u)\mathbf{1}ds,$$

where $\tilde{\boldsymbol{P}}_1(u) = \boldsymbol{P}_1\exp(\boldsymbol{G}tu)$ and $\tilde{\boldsymbol{P}}_2(u) = \exp(\boldsymbol{G}t(1-u))\boldsymbol{P}_2$. By the definition of generator matrix, $\exp(\boldsymbol{G}tu)$ and $\exp(\boldsymbol{G}t(1-u))$ are transition probability matrices from time 0 to time tu and $t(1-u)$ respectively. Hence, both $\tilde{\boldsymbol{P}}_1(u)$ and $\tilde{\boldsymbol{P}}_2(u)$ are transition probability matrices and the summation of each row equals 1. That is, $e_l'\tilde{\boldsymbol{P}}_2(u)\mathbf{1} = e_k'\tilde{\boldsymbol{P}}_2(u)\mathbf{1} = 1$, which implies $\pi'\tilde{\boldsymbol{P}}_1(u)e_k(e_l - e_k)'\tilde{\boldsymbol{P}}_2(u)\mathbf{1} = 0$, and (9) is derived.

Proof of Theorem 1. Let θ_h denote the hth element of θ. It suffices to show that the observed likelihood $L_i(\theta)$ defined in (4) satisfy the regular conditions for maximum likelihood estimator [23]. Specifically, in addition to Condition (C4), we need to show that

$$E_\theta\big[\frac{\partial \log L_i(\theta)}{\partial\theta_h}\big] = 0, \tag{10}$$

$$E_\theta\big[\frac{\partial \log L_i(\theta)}{\partial\theta_h}\frac{\partial \log L_i(\theta)}{\partial\theta_k}\big] = -E_\theta\big[\frac{\partial^2 \log L_i(\theta)}{\partial\theta_h\partial\theta_k}\big], \tag{11}$$

and

$$\Big|\frac{\partial^3}{\partial\theta_h\partial\theta_k\partial\theta_l}\log L_i(\theta)\Big| \leq M_i^{(hkl)}, \tag{12}$$

for some $\tilde{M}_i^{(hkl)}$ with finite expectations.

For the proof of (10), we first note that

$$E_\theta\big[\frac{\partial \log L_i(\theta)}{\partial\theta_h}\big] = \int \frac{\partial}{\partial\theta_h}L_i(\theta)dy_{i1}\ldots dy_{in_i},$$

where

$$\frac{\partial}{\partial\theta_h}L_i(\theta) = [\frac{\partial}{\partial\theta_h}\pi_i'][\prod_{j=1}^{n_i}\boldsymbol{F}_{ij}(\theta)\mathbf{1}] + \sum_{j=1}^{n_i}\pi_i'\{\prod_{v=1}^{j-1}\boldsymbol{F}_{iv}(\theta)\}[\frac{\partial}{\partial\theta_h}\boldsymbol{F}_{ij}(\theta)]\{\prod_{v=j+1}^{n_i}\boldsymbol{F}_{iv}(\theta)\}\mathbf{1}, \tag{13}$$

with $\boldsymbol{F}_{ij} = \exp(\boldsymbol{G}_i\tilde{t}_{ij})\tilde{\boldsymbol{Y}}_{ij}$ and $\tilde{\boldsymbol{Y}}_{ij} = diag\{f(Y_{ij}; X_{ij}, \beta_k, \phi), k \in \{1,\ldots,d\}\}$.

When $\theta_h = \eta_{mk}$ for $m = 1,\ldots,q$ and $k \in \{1,\ldots,d-1\}$, the second term of (13) is 0, and we have that $\frac{\partial}{\partial\eta_{mk}}\pi_i = W_{im}\pi_{ik}(e_k - \pi_i)$. Notice that $\int \tilde{\boldsymbol{Y}}_{ij}dy_{ij} = \boldsymbol{I}$, we then have $\int \boldsymbol{F}_{ij}(\theta)dy_{ij} = \exp(\boldsymbol{G}_i\tilde{t}_{ij})$. Therefore,

$$\begin{aligned}
E_\theta\big[\frac{\partial \log L_i(\theta)}{\partial\eta_{mk}}\big] &= W_{im}\pi_{ik}(e_k - \pi_i)\int\prod_{j=1}^{n_i}\boldsymbol{F}_{ij}(\theta)\mathbf{1}dy_{i1}\ldots dy_{in_i} \\
&= W_{im}\pi_{ik}(e_k - \pi_i)\prod_{j=1}^{n_i}\big[\boldsymbol{F}_{ij}(\theta)dy_{ij}\mathbf{1}\big] \\
&= W_{im}\pi_{ik}(e_k - \pi_i)\exp(\boldsymbol{G}_i t_{in_i})\mathbf{1},
\end{aligned}$$

where $t_{in_i} = \sum_{j=1}^{n_i} \tilde{t}_{ij}$ is the last observation time of subject i. Note that $\exp(\boldsymbol{G}_i t_{in_i})$ is a transition probability matrix whose row summation is 1. Therefore, $(e_k - \pi_i)\exp(\boldsymbol{G}_i t_{in_i})\mathbf{1} = (e_k - \pi_i)\mathbf{1} = 0$, which completes the proof of (10) for $\theta_h = \eta_{mk}$, $m \in \{1, \ldots, q\}$.

When $\theta_h = \beta_{mk}$ for $m \in \{1, \ldots, p\}$, $k \in \{1, \ldots, d\}$ or $\theta_h = \phi_m$ for $m \in \{1, \ldots, r\}$, the first term of (13) is 0 and we have the following:

$$\begin{aligned} E_\theta\Big[\frac{\partial \log L_i(\theta)}{\partial \theta_h}\Big] &= \sum_{j=1}^{n_i} \pi_i' \int \Big\{\prod_{v=1}^{j-1} \boldsymbol{F}_{iv}(\theta)\Big\}\Big[\frac{\partial}{\partial \theta_h}\boldsymbol{F}_{ij}(\theta)\Big]\Big\{\prod_{v=j+1}^{n_i} \boldsymbol{F}_{iv}(\theta)\Big\}\mathbf{1} dy_{i1}\ldots dy_{in_i} \\ &= \sum_{j=1}^{n_i} \pi_i' \Big\{\prod_{v=1}^{j-1} \int \boldsymbol{F}_{iv}(\theta) dy_{iv}\Big\} \int \Big[\frac{\partial}{\partial \theta_h}\boldsymbol{F}_{ij}(\theta)\Big] dy_{ij} \Big\{\prod_{v=j+1}^{n_i} \int \boldsymbol{F}_{iv}(\theta) dy_{iv}\Big\}\mathbf{1} \\ &= \sum_{j=1}^{n_i} \pi_i' \exp(\boldsymbol{G}_i t_{ij}) E_\theta\Big[\frac{\partial}{\partial \theta_h}\log(\tilde{\boldsymbol{Y}}_{ij})\Big]\exp(\boldsymbol{G}_i(t_{in_i} - t_{ij}))\mathbf{1}. \end{aligned}$$

By Condition (C2), we have that $E_\theta[\frac{\partial}{\partial \theta_h}\log(\tilde{\boldsymbol{Y}}_{ij})] = 0$, which implies (10) holds for $\theta_h = \beta_{mk}$, $m \in \{1, \ldots, p\}$, $k \in \{1, \ldots, d\}$, and $\theta_h = \phi_m$, $m \in \{1, \ldots, r\}$.

Finally, when $\theta_h = \gamma_{klm}$ for $m \in \{1, \ldots, s\}$, $k, l \in \{1, \ldots, d\}$, $k \neq l$, the first term of (13) is 0, similar to the aforementioned discussion, we have

$$E_\theta\Big[\frac{\partial \log L_i(\theta)}{\partial \gamma_{klm}}\Big] = \sum_{j=1}^{n_i} \pi_i' \exp(\boldsymbol{G}_i t_{i,j-1})\Big[\frac{\partial}{\partial \gamma_{klm}}\exp(\boldsymbol{G}_i \tilde{t}_{ij})\Big]\exp(\boldsymbol{G}_i(t_{in_i} - t_{ij}))\mathbf{1}.$$

Note that

$$\frac{\partial}{\partial \gamma_{klm}}\exp(\boldsymbol{G}_i \tilde{t}_{ij}) = \frac{\partial g_{ikl}}{\partial \gamma_{klm}}\frac{\partial}{\partial g_{ikl}}\exp(\boldsymbol{G}_i \tilde{t}_{ij}),$$

In addition, both $\exp(\boldsymbol{G}_i t_{i,j-1})$ and $\exp(\boldsymbol{G}_i(t_{in_i} - t_{ij}))$ are transition probability matrices. By Lemma A.3, we have that (10) holds for $\theta_h = \gamma_{kl}$, $k, l \in \{1, \ldots, d\}$, $k \neq l$. Thus the proof for that (10) holds for all components of θ is completed.

For the proof of (11), taking derivative of (10) with respect to θ_k, we obtain that

$$\frac{\partial}{\partial \theta_k}\Big[\int L_i(\theta)\frac{\partial \log L_i(\theta)}{\partial \theta_h} dy_{i1}\ldots dy_{in_i}\Big] = 0.$$

By the boundedness of the derivatives, we can change the order of derivation and integration, thus we obtain

$$E_\theta\Big[\frac{\partial \log L_i(\theta)}{\partial \theta_h}\frac{\partial \log L_i(\theta)}{\partial \theta_k}\Big] + E_\theta\Big[\frac{\partial^2 \log L_i(\theta)}{\partial \theta_h \partial \theta_k}\Big] = 0,$$

which implies Eq. (11).

The inequality (12) can be easily derived by Condition (C3) and the boundedness of derivatives with respect to η_{mk} and γ_{klm}. Thus, we have completed the proof of Theorem 1.

Proof of Theorem 2. Without loss of generality, we omit the subscripts of $\boldsymbol{G}$ and take $k_0 = k_1 = d$ and $k_2 = d + 1$. For simplicity, we only give the proof for one sample with one observation. It can be easily extended to a general case. We accept the notations introduced in Lemmas A.1 and A.2. Note that if $\boldsymbol{G}$ and $\boldsymbol{G}^{(new)}$ satisfy Condition (iii), then, for any $t > 0$, $\boldsymbol{G}t$ and $\boldsymbol{G}^{(new)}t$ also satisfy that condition. Therefore, by Lemma A.2, $\exp\{\boldsymbol{G}t\} \prec \exp\{\boldsymbol{G}^{(new)}t\}$. Let $\pi = (\pi'_{(d-1)}, \pi_d)'$ be any initial distribution for d states with $\pi_{(d-1)} \in R^{d-1}$ and split the last probability π_d into two parts: $\pi_d^{(1)} + \pi_d^{(2)} = \pi_d$. We recall that $\tilde{\boldsymbol{Y}}_{11}$ is the $d \times d$ diagonal matrix consisted of $\{f(Y_1(t_{11}); X_1(t_{11}), \beta_k, \phi), k \in \{1, \ldots, d\}\}$ and denote $\tilde{\boldsymbol{Y}}_{11}^{(d+1)}$ for the $(d+1)\times(d+1)$ diagonal matrix with the first d diagonal elements the same as $\tilde{\boldsymbol{Y}}_{11}$ and the last diagonal element being $f(Y_1(t_{11}); X_1(t_{11}), \beta_d, \phi)$. It suffices to show that, for any two related transition probability matrices $\boldsymbol{B} \prec \boldsymbol{D}$, it holds that $L^{(d)} = L^{(d+1)}$ where

$$L^{(d)} = \pi' B \tilde{\boldsymbol{Y}}_{11}\mathbf{1}, \qquad L^{(d+1)} = (\pi'_{(d-1)}, \pi_d^{(1)}, \pi_d^{(2)})\boldsymbol{D}\tilde{\boldsymbol{Y}}_{11}^{(d+1)}\mathbf{1}.$$

This can be verified by the block matrix approach used in the proof of Lemma A.1.

Proof of Theorem 3. Let $\theta^{(d_0)}$ be the parameter corresponding to d_0 states and $\theta_0^{(d_0)}$ be the true parameter. Let $\hat{\theta}^{(d_0)}$ be the maximum of $\tilde{L}(\theta^{(d_0)})$. Note that d_0 is unknown in practice, hence $\hat{\theta}^{(d_0)}$ cannot be derived in reality. With the result of Theorem 1, following the lines of [11], we can show that $\sqrt{n}(\hat{\theta}^{(d_0)} - \theta_0^{(d_0)})$ is asymptotically normal. We extend the notation σ_k defined in Section 4.2 to evolve θ. Note that both of $\hat{\theta}^{(d_0)}$ and $\theta_0^{(d_0)}$ have $d_0 < d$ ordered distinct states. With the result of Theorem 2, we can extend the number of states while preserving the likelihood unchanged by duplicating existing states.

Let $\hat{\theta}^{(d)}$ and $\theta_0^{(d)}$ be one version of $\hat{\theta}^{(d_0)}$ and $\theta_0^{(d_0)}$ in the parameter space with d states satisfying $\hat{\theta}^{(d)} - \theta_0^{(d)} = O_p(n^{-1/2})$. Then, $\hat{\theta}^{(d)}$ and $\theta_0^{(d)}$ must have the same state structures with d_0 distinct states. Let $s_l, l \in \{1, \ldots, d_0\}$ be the indexes of the first one of each distinct states, then, for all $k \neq s_l - 1, l \in \{2, \ldots, d_0\}$ we have that $\sigma_k(\theta_0^{(d)}) = \sigma_k(\hat{\theta}^{(d)}) = 0$, where $\sigma_k(\theta)$ is an extension of the notation σ_k defined in Section 4.2 to evolve θ. Therefore, for any given version of $\theta_0^{(d)}$ we have that

$$\hat{\theta}^{(d)} = \underset{\sigma_k(\theta)=0, k\neq s_l-1, l\in\{2,\ldots,d_0\}}{\text{Argmax}} \tilde{L}(\theta) \tag{14}$$

On the other hand, for any θ satisfying $\theta - \theta_0^{(d)} = O_p(n^{-1/2})$ and $\epsilon > 0$, we show that, with probability tending to 1,

$$\tilde{L}(\theta) < \tilde{L}(\hat{\theta}^{(d)}), \quad \text{If} \max_{k\neq s_l-1, l\in\{2,\ldots,d_0\}} \sigma_k(\theta) \geq \epsilon n^{-1/2}. \tag{15}$$

Thus we conclude that $\hat{\theta}^{(d)}$ is a local maximum of $\tilde{L}(\theta)$ by combining (14) and (15). To prove (15), by Taylor's expression, we have

$$L(\theta) - L(\hat{\theta}^{(d)}) = \sum_k \frac{\partial L(\theta_0^{(d)})}{\partial \theta_k}(\theta_k - \hat{\theta}_k^{(d)}) + \sum_{k,l} \frac{\partial^2 L(\theta^*)}{\partial \theta_k \partial \theta_l}(\theta_k - \hat{\theta}_k^{(d)})(\theta_l - \hat{\theta}_l^{(d)}),$$

where θ^* lies between θ and $\hat{\theta}^{(d)}$. Similar to the derivation of (10), we can show that $E_{\theta_0^{(d)}} \frac{\partial \log\{L_i(\theta_0^{(d)})\}}{\partial \theta_k} = 0$. Therefore, $\frac{\partial L(\theta_0^{(d)})}{\partial \theta_k} = O_p(n^{1/2})$. It is easy to obtain that $\frac{\partial^2 \tilde{L}(\theta^*)}{\partial \delta_k \partial \delta_l} = O_p(n)$. Since both $\theta - \theta_0^{(d)}$ and $\hat{\theta}^{(d)} - \theta_0^{(d)}$ are $O_p(n^{-1/2})$, we have that

$$L(\theta) - L(\hat{\theta}^{(d)}) = O_p(1). \tag{16}$$

For the penalty term, note that for any $k = s_l - 1, l \in \{2, \ldots, d_0\}$, both $\sigma_k(\theta)$ and $\sigma_k(\hat{\theta}^{(d)})$ are bounded away from 0 and therefore the condition $\lambda_N \to 0$ implies that $p_{\lambda_N}(\sigma_k(\theta)) = p_{\lambda_N}(\sigma_k(\hat{\theta}^{(d)}))$ since $\sigma_k(\theta) - \sigma_k(\hat{\theta}^{(d)}) = O_p(n^{-1/2})$ and the property of the SCAD penalty. While for those $k \neq s_l, l \in \{2, \ldots, d_0\}$, by assumption, we have $\sigma_k(\hat{\theta}^{(d)}) = 0$ and $\sigma_k(\theta) = O_p(n^{-1/2})$ which implies $p_{\lambda_N}(\sigma_k(\hat{\theta}^{(d)})) = 0$ and $p_{\lambda_N}(\sigma_k(\theta)) = \lambda_N \sigma_k(\theta)$. Therefore,

$$N \sum_{k=1}^{d-1} p_{\lambda_N}(\sigma_k(\theta)) - N \sum_{k=1}^{d-1} p_{\lambda_N}(\hat{\sigma}_k(\hat{\theta}^{(d)})) = N\lambda_N \sum_{k\neq s_l, l=2,\ldots,d_0} \sigma_k(\theta).$$

If $\max_{k\neq s_l-1, l\in\{2,\ldots,d_0\}} \sigma_k(\theta) \geq \epsilon n^{-1/2}$, we obtain that $N\lambda_N \sum_{k\neq s_l, l\in\{2,\ldots,d_0\}} \sigma_k(\theta) \geq N n^{-1/2} \lambda_N \epsilon$, which goes to infinity by the condition $\sqrt{n}\lambda_N \to \infty$ and $N \geq n$. Combining this with (16), we complete the proof of (15).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2020.104646.

References

[1] H.F. Albert, M.E. McFarland, M. E, J.A. Frank, Time series for modleing counts from a relapsing-remitting disease: application to modelling disease activity in multiple sclerosis, Stat. Med. 13 (1994) 453–466.
[2] R.J. Altman, Mixed hidden Markov models: An extension of the hiddenMarkov model to the longitudinal data setting, J. Amer. Statist. Assoc. 102 (2007) 201–210.
[3] G. Baetschmann, R. Winkelmann, Modeling zero-inflated count data when exposure varies: with an application to tumor counts, Biometrical J. 55 (2007) 679–686.
[4] N.T.J. Bailey, The Elements of Stochastic Processes, Wiley, New York, 1964.
[5] F. Bartolucci, A. Farcomeni, A shared-parameter continuous-time hidden Markov and survival model for longitudinal data with informative dropout, Stat. Med. 38 (2019) 1056–1073.
[6] F. Bartolucci, A. Farcomeni, F. Pennoni, Latent Markov Models for Longitudinal Data, Chapman & Hall/CRC, Taylor and Francis Group, 2013.
[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (2011) 1–122.
[8] A. Bureau, S. Shiboski, J.P. Hughes, Applications of continuous time hidden markov models to the study of misclassified disease outcomes, Stat. Med. 22 (2003) 441–462.
[9] O. Cappé, E. Moulines, T. Rydén, Inference in Hidden Markov Models, Springer, New York, 2005.
[10] J. Chen, A. Khalili, Order selection in finite mixture models with a nonsmooth penalty, J. Amer. Statist. Assoc. 103 (2008) 1674–1683.
[11] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[12] A. Farcomeni, Penalized estimation in latent Markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency, Biometrical J. 59 (2017) 1035–1046.
[13] P. Fearnhead, C. Sherlock, An exact Gibbs sampler for the Markov-modulated Poisson process, J. R. Stat. Soc. Ser. B Stat. Metholol. 68 (2006) 767–784.
[14] J. Gallier, Logarithms and square roots of real matrices, arXiv:0805.0245.
[15] Y. Guédon, Estimating hidden semi-Markov chains from discrete sequences, J. Comput. Graph Stat. 12 (2003) 604–639.
[16] C. Guihenneuc-Jouyaux, S. Richardson, I.M. Longini, Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline, Biometrics 56 (2000) 733–741.

[17] Y. Hung, Y. Wang, V. Zarnitsyna, C. Zhu, C.J. Wu, Hidden Markov models with applications in cell adhesion experiments, J. Amer. Statist. Assoc. 108 (2013) 1469–1479.
[18] E.H. Ip, Q. Zhang, W.J. Rejeski, T.B. Harris, S. Kritchevsky, Partially ordered mixed hidden Markov model for the disablement process of older adults, J. Amer. Statist. Assoc. 108 (2013) 370–384.
[19] K. Kang, J. Cai, X. Song, H. Zhu, Bayesian hidden Markov models for delineating the pathology of Alzheimer's disease, Stat. Methods Med. Res. 28 (2019) 2112–2124.
[20] R. Langrock, I.L. MacDonald, W. Zucchini, Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models, J. Empir. Finance 19 (2012) 147–161.
[21] R. Langrock, W. Zucchini, Hidden Markov models with arbitrary state dwell-time distributions, Comput. Statist. Data Anal. 55 (2011) 715–724.
[22] G.J. Lastman, N.K. Sinha, Infinite series for logarithm of matrix, applied to identification of linear continuous-time multivariable systems from discrete-time models, Electron. Lett. 27 (1991) 1468–1470.
[23] E.L. Lehmann, Theory of Point Estimation, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1983.
[24] Y. Liang, W. Lu, Z. Ying, Joint modeling and analysis of longitudinal data with informative observation times, Biometrics 65 (2009) 377–384.
[25] Y.Y. Liu, S. Li, F. Li, L. Song, J.M. Rehg, Efficient learning of continuous-time hidden Markov models for disease progression, Adv. Neural Inf. Process. Syst. (2015) 3600–3608.
[26] I.L. MacDonald, W. Zucchini, Hidden Markov Models and Other Models for Discrete-Valued Time Series, Chapman & Hall, London, 1997.
[27] R.J. MacKAY, Estimating the order of a hidden Markov model, Canad. J. Statist. 30 (2002) 573–589.
[28] A. Maruotti, Mixed hidden Markov models for longitudinal data: an overview, Int. Stat. Rev. 79 (2011) 427–454.
[29] C. Moler, C. van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, SIAM Rev. 45 (2003) 3–49.
[30] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 259–286.
[31] S.M. Ross, Stochastic Processes, John Wiley & Sons, New York, 1996.
[32] S. Scott, G. James, C. Sugar, Hidden Markov models for longitudinal comparisons, J. Amer. Statist. Assoc. 100 (2005) 359–369.
[33] S.L. Scott, P. Smyth, The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modeling, Bayesian Stat. 7 (2003) 1–7.
[34] X. Song, Y. Xia, H. Zhu, Hidden Markov latent variable models with multivariate longitudinal data, Biometrics 73 (2017) 313–323.
[35] J. Sun, D.H. Park, L. Sun, X. Zhao, Semiparametric regression analysis of longitudinal data with informative observation times, J. Amer. Statist. Assoc. 100 (2005) 882–889.
[36] L. Sun, X. Song, J. Zhou, Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times, J. Statist. Plann. Inference 141 (2011) 2902–2919.
[37] J. Sun, L. Sun, D. Liu, Regression analysis of longitudinal data in the presence of informative observation and censoring times, J. Amer. Statist. Assoc. 102 (2007) 1397–1406.
[38] C. Van Loan, Computing integrals involving the matrix exponential, IEEE Trans. Automat. Control 23 (1978) 395–404.
[39] R.M. Wilcox, Exponential operators and parameter differentiation in quantum physics, J. Math. Phys. 8 (1967) 962–982.
[40] X. Zhou, K. Kang, X. Song, Two-part hidden Markov models for semicontinuous longitudinal data with nonignorable missing covariates, Stat. Med. 39 (2020) 1801–1816.