# Journal Pre-proof


Order selection for regression-based hidden Markov model

Yiqi Lin, Xinyuan Song

PII: S0047-259X(22)00070-7
DOI: https://doi.org/10.1016/j.jmva.2022.105061
Reference: YJMVA 105061

To appear in: *Journal of Multivariate Analysis*

Received date : 17 September 2021
Revised date : 31 May 2022
Accepted date : 1 June 2022

Please cite this article as: Y. Lin and X. Song, Order selection for regression-based hidden Markov model, *Journal of Multivariate Analysis* (2022), doi: https://doi.org/10.1016/j.jmva.2022.105061.

# Order selection for regression-based hidden Markov model

Yiqi, Lin[a], Xinyuan, Song[a,*]

[a]*Department of Statistics, The Chinese University of Hong Kong, Hong Kong*

**Abstract**

Hidden Markov models (HMMs) describe the relationship between two stochastic processes: an observed process and an unobservable finite-state transition process. Owing to their modeling dynamic heterogeneity, HMMs are widely used to analyze heterogeneous longitudinal data. Traditional HMMs frequently assume that the number of hidden states (i.e., the order of HMM) is a constant and should be specified prior to analysis. This assumption is unrealistic and restrictive in many applications. In this study, we consider regression-based hidden Markov model (RHMM) while allowing the number of hidden states to be unknown and determined by the data. We propose a novel likelihood-based double penalized method, along with an efficient expectation-conditional maximization with iterative thresholding-based descent (ECM–ITD) algorithm, to perform order selection in the context of RHMM. An extended Group-Sort-Fuse procedure is proposed to rank the regression coefficients and impose penalties on the discrepancy of adjacent coefficients. The order selection consistency and convergence of the ECM–ITD algorithm are established under mild conditions. Simulation studies are conducted to evaluate the empirical performance of the proposed method. An application of the proposed methodology to a real-life study on Alzheimer's disease is presented.

*Keywords:* ECM–ITD algorithm, Group-Sort-Fuse procedure, hidden Markov model, longitudinal data, order selection.
*2022 MSC:* Primary 62M05, Secondary 62J07

## 1. Introduction

Hidden Markov model (HMM) is a common statistical tool for simultaneously analyzing a longitudinal observation process and its dynamic transition process. HMMs have been widely applied to many research fields, such as computational biology, medical studies, natural language processing, and time series forecasting, in the past two decades. Examples of the HMMs applications and their variants include gene finding [33] and protein sequence alignment [11] in computational biology; automatic speech recognition [18, 41] and activity recognition [38] in engineering; multiple sclerosis, late-life disability, cocaine addiction, and Alzheimer's disease (AD) progression and prevention [3, 16, 20, 36] in medical studies.

Majority of existing HMMs either assume that the number of hidden states (i.e., order of HMM) is a prespecified constant or determine the order using criterion-based methods, such as AIC [2] and BIC [32]. Although these information criteria provide a general framework for the component/order selection of mixture-type models and have been successfully applied to many substantive studies [16, 22, 36], their use in the context of HMMs has not been theoretically justified [25]. Alternative methods for order selection of HMMs include techniques developed under Bayesian framework, such as variational Bayes approach [5] and reversible jump Markov chain Monte Carlo algorithm [13, 23]. However, such Bayesian order selection methods still suffer from lack of theoretical guarantee.

A few studies have investigated the theoretical properties of order selection in the context of mixture-type models. For example, MacKAY [26] introduced the penalized minimum-distance method, which imposes a penalty only on a small proportions of states, yielding a consistent order estimate of HMM. However, Chen and Khalili [8] argued that

*Corresponding author. Email address: xysong@sta.cuhk.edu.hk

the second type of overfitting exists in mixture-type models because some component densities may be close to each other. Such overfitting is undesirable because it also reduces order selection performance. To overcome this limitation, the researchers proposed a double penalized likelihood method to simultaneously prevent the two types of overfitting. Hung et al. [15] further extended the double penalized method to conduct order selection in the context of Gaussian HMM. Nevertheless, the preceding works mainly focused on finite mixture models or non-regression-based HMMs. Zhou et al. [43] proposed a modified penalty method to conduct order selection in continuous-time HMMs. To the best of our knowledge, order selection for regression-based HMM (RHMM) with state-specific covariate effects has been a largely underexplored domain.

In this study, we propose a novel double penalized procedure to conduct order selection in the context of RHMM. The proposed model consists of two parts: a conditional regression model to assess the associations between the longitudinal response and multiple covariates, and a transition model to formulate transition from one state to another. We introduce two penalties terms into the likelihood function, namely, lower bound on mixing probabilities and smoothly clipped absolute deviations (SCAD) penalty [12] to prevent far closed states. The Group-Sort-Fuse (GSF) procedure proposed by [27] is extended to rank the multidimensional regression coefficients in maximizing the penalized likelihood to perform order selection and obtain the maximum penalized likelihood estimator (MPLE). We also establish the order selection consistency under mild regularity and identifibiality conditions. An efficient expectation-conditional maximization with iterative thresholding-based descent (ECM–ITD) algorithm is developed to facilitate order selection and parameter estimation, and its convergence is guaranteed by theoretical results.

The rest of this article is organized as follows. Section 2 describes the proposed RHMM and associated order selection procedure. The consistency of order selection is also established. Section 3 develops the ECM–ITD algorithm for parameter estimation. Some related computational issues and tuning strategies are also discussed. Section 4 presents simulation studies to examine the empirical performance of the proposed method. In section 5, an application to Alzheimer's Disease Neuroimaging Initiative (ADNI) study is presented. Section 6 concludes the paper. Technical details are provided in the Appendix, and additional information and numerical results are presented in Online Supplementary Material.

## 2. Regression-based Hidden Markov Model

### *2.1. Model Description*

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$, where $\boldsymbol{Y}_i = \{y_{it}\}$, $t \in \{1, \ldots, T\}$ and $y_{it}$ be the response of subject $i$ at time $t$; $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_n)$, where $\boldsymbol{S}_i = \{S_{it}\}$, $t \in \{1, \ldots, T\}$, is a set of hidden states associated with $y_{it}$, and $S_{it}$ is assumed to be a finite-state stationary Markov chain taking values in $\{1, \ldots, K\}$. The transition between different states can be described by a homogeneous transition matrix $P = [P_{rs}]_{K\times K}$ with $P_{rs} = P(S_{it} = s|S_{i,t-1} = r)$ for $\forall i$ and $t \in \{2, \ldots, T\}$ and stationary probability $\pi_r$, where $r, s \in \{1, \ldots, K\}$. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ be the set of covariates, where $\boldsymbol{X}_i = \{\boldsymbol{x}_{it}\}$, $t \in \{1, \ldots, T\}$, and $\boldsymbol{x}_{it}$ is a $(q + 1) \times 1$ vector of covariates for subject $i$ at time $t$. Conditional on hidden state $S_{it} = k$ and covariates $\boldsymbol{x}_{it}$, a generalized linear model for response $y_{it}$ is considered as follows:

$$\begin{aligned} f(y_{it}|S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}_k) &= \exp\left\{(y_{it}\theta_{itk} - b(\theta_{itk}))/a(\phi) + c(y_{it}, \phi)\right\}, \\ \theta_{itk} &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k, \end{aligned} \tag{1}$$

where $f(y_{it}|\cdot)$ denotes the probability mass/density function of $y_{it}$; $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are some specific functions; $\theta_{itk}$ is the state-specific canonical parameter; $\phi$ is a dispersion parameter; and $\boldsymbol{\beta_k}$ is a vector of state-specific regression coefficients. The conditional mean and variance of $y_{it}$ given $S_{it} = k$ and $\boldsymbol{x}_{it}$ are $E(y_{it}|S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}_k) = b'(\theta_{itk})$ and $\text{var}(y_{it}|S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}_k) = b''(\theta_{itk})/a(\phi)$. In this study, we assume that $a(\phi) = 1$ for simplicity, an extension to the case where $a(\phi) \neq 1$ is straightforward.

In the proposed RHMM, the number of hidden states (i.e., order of HMM) is unknown and must be determined by the data. For ease of exposition, $K$ and $K_0$ are denoted as the upper bound and true value of the order, respectively. Let $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_K; P_{11}, P_{12}, \ldots, P_{KK}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$. Then, the probability mass/density function of $\boldsymbol{Y}_i$ can be written as

$$F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}) = \sum_{S_{i1}=1}^{K} \cdots \sum_{S_{iT}=1}^{K} \left[\prod_{t=1}^{T} \left[f\left(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}}\right)\right] \pi_{S_{i1}} P_{S_{i1}S_{i2}} \times \cdots \times P_{S_{i,T-1}S_{iT}}\right]. \tag{2}$$

The conditional regression model (1) can accommodate responses of different distributions, including familiar binomial, Poisson, and normal distributions. Compared with a Gaussian HMM without a conditional regression model [15], the proposed RHMM enables us to examine the state-specific effects of potential covariates on the response of interest. On the basis of Equation (2), the proposed RHMM can be regarded as a dynamic finite mixture of regression models, where the covariate effects are allowed to vary across states over time.

### *2.2. Order Selection via Extended GSF Procedure*

Determining the number of hidden states (order selection) is an important issue in the analysis of HMMs. Only a few states may cause a bad fit, whereas a large number of states may lead to data overfitting. When the number of hidden states has a specific meaning, for example, the number of phases of AD progression, an inaccurate order estimate leads to a fatal misleading inference.

Suppose $K_0$ is the true number of hidden states. A natural way to estimate the order of RHMM is the MLE of overfitted log-likelihood with the upper bound of order $K$ ($K \geq K_0$):

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log F\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}\right). \tag{3}$$

However, the overfitted MLE leads to an inconsistent estimate of $K_0$ [8, 15, 27]. The overfitting of MLE is of two types, namely, (1) near-zero values of mixing probability and (2) the densities of some components are close to each other. [8] proposed a double penalized method in the context of a finite mixture model to avoid the aforementioned overfitting problem. The first penalty prevents type (1) overfitting by imposing a lower bound on mixing probabilities $\sum_{k=1}^{K} \log \pi_k$, whereas the second penalty $p_{\lambda_n}(\cdot)$ circumvents type (2) overfitting by shrinking to zero the discrepancy between the fitted atoms that are close to each other. This method has been successfully applied to Gaussian HMM [15] and multivariate finite mixture model [27]. As the preceding developments focused either on cross-sectional mixture models or non-regression-based HMMs, they are not directly applicable to the present study. In the following, we propose a new double penalized method to conduct simultaneous order selection and parameter estimation for the proposed RHMM.

The double penalized log-likelihood can be written as follows:

$$\tilde{\ell}_n(\boldsymbol{\Psi}) = \ell_n(\boldsymbol{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - n \sum_{k=2}^{K} p_{\lambda_n}\left(\|\boldsymbol{\eta_k}\|_2\right), \tag{4}$$

where $C_K$ is a tuning parameter, $p_{\lambda_n}(\cdot)$ is a penalty function, and $\|\boldsymbol{\eta}_k\|_2 = \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2$, which will be clarified in the subsequent section. Numerous penalty functions are available in the literature. For example, Tibshirani [37] introduced the least absolute shrinkage and selection operator (LASSO) penalty in the context of linear regression model. Thereafter, Fan and Li [12] proposed a SCAD penalty that retains the penalization rate of LASSO for small values but relaxes the penalty for large ones. The minimax concave penalty (MCP) [42] and adaptive LASSO penalty [44] also share a similar asymptotic property of SCAD. In this study, we consider the SCAD penalty because of its popularity and desirable asymptotic properties. The proposed procedure can be extended by incorporating the MCP and adaptive LASSO penalties without much difficulty. The specific form of the SCAD penalty can be characterized by its derivative as follows:

$$p'_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \leq \lambda_n) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda_n) \right\}, \tag{5}$$

where $a > 2$ and $\lambda_n$ are tuning parameters.

**Remark 1.** We want to highlight the necessity of double penalties in detecting redundant states. The existence of redundant states can cause near-zero mixing probabilities or near-identical parameter values in certain states. Thus, we introduce the first and second penalties to avoid near-zero mixing probabilities and shrink the differences of near-identical parameter values to zero, respectively, to simultaneously eliminate the redundant signals appearing in the mixing probability and model parameters. Only using either penalty cannot achieve this purpose.

Notably, the double penalized methods developed by [8, 15] focused on a special case where the mixture components/hidden states are determined by a one-dimensional parameter. Their methods introduced the first penalty on the mixing/initial probabilities and the second penalty to penalize the discrepancy of sorted atoms of the mixed components in accordance with the natural ranking of the one-dimensional parameters. However, their methods cannot be directly applied to the present study because the multidimensional state-specific regression coefficients have no such natural order. A possible solution to the problem is to penalize the norm of the pairwise discrepancy of different regression coefficients to avoid sorting the multidimensional parameters. [27] investigated this penalization method in the context of finite mixture models with multidimensional parameters and found that the sensitivity of estimator to upper bound $K$ increases when naively applying penalty to all $\binom{K}{2}$ pairwise differences. Thus, they proposed a so-called GSF procedure to sort the multidimensional parameters and then merely penalize $K-1$ consecutive differences. Inspired by this idea, we develop an extended GSF procedure for the regression coefficients of different hidden states in the proposed RHMM as follows.

We first adapt the definitions 1 of [27] to our model up to some changes of notations.

**Definition 1.** Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \Theta \subseteq \mathbb{R}^d$, and $\mathcal{P}_H = \{C_1, \ldots, C_H\}$ be a partition of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$, for any integer $1 \leq H \leq K$. Suppose

$$\max_{\boldsymbol{\beta}_i, \boldsymbol{\beta}_j \in C_h} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2 < \min_{\boldsymbol{\beta}_i \in C_h, \boldsymbol{\beta}_l \notin C_h} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_l\right\|_2, \quad h \in \{1, \ldots, H\}. \tag{6}$$

Then, each set $C_h$ is an atom cluster and $\mathcal{P}_H$ is a cluster partition.

This definition indicates that the largest distance within an atom cluster is strictly smaller than any distance between two atoms not in the same cluster. Intuitively, we expect all estimated $\boldsymbol{\beta}_k$ to fall into a reasonable cluster partition $\mathcal{P}_{K_0}$, within which each $C_h$ contains at least one $\boldsymbol{\beta}_k$. Therefore, we can merge all the estimated $\boldsymbol{\beta}_k$ in a $C_h$ to form one estimate of $\hat{\boldsymbol{\beta}}_h$ for $h \in \{1, \ldots, K_0\}$. Assuming the existence of the reasonable cluster partition $\mathcal{P}_{K_0}$, SCAD penalty is desirable to merge the elements in each $C_h$ and preserve the discrepancy of each pair in $\mathcal{P}_{K_0}$. Moreover, a cluster ordering that provides such reasonable $K-1$ pairs of $\boldsymbol{\beta}_k$ is required to efficiently implement the aforementioned penalization procedure.

Recently, [43] proposed a cluster ordering procedure, which coincides with that proposed by [27] for finite mixture models with multidimensional parameters, to perform order selection for continuous-time HMMs. Motivated by [43], a cluster ordering procedure can be defined as follows:

**Definition 2.** For $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \Theta \subseteq \mathbb{R}^d$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$, $\alpha_{\boldsymbol{\beta}} = \left\{\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(K)}\right\}$ is called a cluster ordering procedure if

$$\boldsymbol{\beta}_{(k)} = \underset{j \notin \{\boldsymbol{\beta}_{(i)}: 1 \leq i \leq k-1\}}{\operatorname{argmin}} \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{(k-1)}\right\|_2, \quad k \in \{2, \ldots, K\}, \tag{7}$$

and $\boldsymbol{\beta}_{(1)} = \underset{k=1,2,\ldots,K}{\operatorname{argmax}} \|\boldsymbol{\beta}_k\|_2$.

The procedure defined in (7) picks the element of $\boldsymbol{\beta}$ with the largest norm as the first, then recursively chooses one closest to the previously picked elements from the remaining. Notably, such a procedure satisfies a so-called atom property mentioned by [27]. That is, $\alpha_{\boldsymbol{\beta}}$ maps the elements in each $C_h$ into a set of consecutive integers.

The cluster ordering procedure $\alpha_{\boldsymbol{\beta}}$ induces a clustered ordered set $\{\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(K)}\}$, so that $\boldsymbol{\eta}_k$ in the penalized log-likelihood $\tilde{l}_n(\boldsymbol{\Psi})$ in Equation (4) can be clarified as $\boldsymbol{\eta}_k = \boldsymbol{\beta}_{(k+1)} - \boldsymbol{\beta}_{(k)}$, $k \in \{1, \ldots, K-1\}$. Then, we can conduct order selection through the extended GSF procedure. Denote $\hat{\boldsymbol{\Psi}}_n = \operatorname{argmax} \tilde{l}_n(\boldsymbol{\Psi})$ as the MPLE of $\boldsymbol{\Psi}$. Then,

$$\hat{K}_n = \text{number of distinct values of} \left\{\hat{\boldsymbol{\beta}}_{(k)}\right\}, k \in \{1, \ldots, K\} \tag{8}$$

is an estimator of true order $K_0$, and we show that $\hat{K}_n$ converges to $K_0$ in probability in the subsequent section. Intuitively, a consistent estimate of the order can be obtained on the two sides. One is that the lower bound penalty $\sum_{k=1}^{K} \log \pi_k$ ensures the existence of cluster partition $\mathcal{P}_{K_0}$ of $\left\{\hat{\boldsymbol{\beta}}_{(1)}, \ldots, \hat{\boldsymbol{\beta}}_{(K)}\right\}$ in an asymptotic manner. Another is that the extended GSF procedure tends to maximize the penalties on the discrepancies of $\hat{\boldsymbol{\beta}}_{(k)}$ in the same cluster and minimize the penalties on the discrepancies of $\hat{\boldsymbol{\beta}}_{(k)}$ in different clusters. The SCAD penalty further ensures $\hat{\boldsymbol{\beta}}_{(k)}$ in the same cluster to be merged into a unique one asymptotically. Given the preceding two-side reasons, the goal of consistent order selection in RHMM can be achieved.

**Remark 2.** The cluster ordering procedure described above is based on [27, 43]. In particular, Manole and Khalili [27] compared the ordering procedure with a naive method that penalizes all $\binom{K}{2}$ pairwise distances in a finite mixture model. They demonstrated that the ordering procedure performs similarly to the naive approach when $K_0$ is small but considerably outperforms the naive one in order selection and computing time when $K_0$ becomes large. Thus, using such an ordering procedure in the proposed RHMM can reduce the unnecessary computation of penalizing pairwise distances of $\boldsymbol{\beta}_k$s.

### *2.3. Asymptotic Properties of Order Selection*

Through a slight abuse of notation, we use $\boldsymbol{\beta}_k$ in the following instead of $\boldsymbol{\beta}_{(k)}$ to denote the $k$th element of $\boldsymbol{\alpha}_{\boldsymbol{\beta}}$ and temporarily suppress subscript $i$ for notation simplicity. Then, the density of $\boldsymbol{Y}$ can be rewritten as follows:

$$F(\boldsymbol{Y};G)=\int F(\boldsymbol{Y};\boldsymbol{X},\boldsymbol{\Psi})\,\mathrm{d}G(\boldsymbol{\Psi})=\sum_{S_1=1}^{K}\cdots\sum_{S_T=1}^{K}\left[f\left(\boldsymbol{Y};\boldsymbol{X},\boldsymbol{\beta}^S\right)\right]\pi_{S_1}P_{S_1S_2}\times\cdots\times P_{S_{T-1}S_T}, \tag{9}$$

where $G(\boldsymbol{\Psi})=\sum_{S_1=1}^{K}\cdots\sum_{S_T=1}^{K}\pi_{S_1}P_{S_1S_2}\times\cdots\times P_{S_{T-1}S_T}\delta(\boldsymbol{\beta}^S)$ is the mixing measure with $\boldsymbol{\beta}^S=\left\{\boldsymbol{\beta}_{S_1},\ldots,\boldsymbol{\beta}_{S_T}\right\}$ and $\delta(\cdot)$ stands for Dirac measure. Accordingly, we denote

$$\hat{G}_n(\boldsymbol{\beta}^S)=\sum_{S_1=1}^{K}\cdots\sum_{S_T=1}^{K}\hat{\pi}_{S_1}\hat{P}_{S_1S_2}\times\cdots\times\hat{P}_{S_{T-1}S_T}\delta(\hat{\boldsymbol{\beta}}^S) \tag{10}$$

as the MPLE of $\tilde{l}_n$ in (4).

We denote by $\boldsymbol{\beta}_{0k}$ the true value of $\boldsymbol{\beta}_k$, for $k\in\{1,\ldots,K\}$ and let $\boldsymbol{\beta}_0=\{\boldsymbol{\beta}_{01},\ldots,\boldsymbol{\beta}_{0K_0}\}$. Inspired by the idea of decision boundary of K-means and in line with [27], we define Voronoi diagram $\hat{\mathcal{V}}_k$ for true $\boldsymbol{\beta}_{0k}$ as

$$\hat{\mathcal{V}}_k=\left\{\hat{\boldsymbol{\beta}}_j:\left\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_{0k}\right\|_2\le\left\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_{0l}\right\|_2,\forall l\ne k\right\},\ k\in\{1,\ldots,K_0\}, \tag{11}$$

to represent the sets of $\hat{\boldsymbol{\beta}}$ closest to $\boldsymbol{\beta}_{0k}$. We define the index set $\hat{\mathcal{I}}_k$ as

$$\hat{\mathcal{I}}_k=\left\{1\le j\le K:\hat{\boldsymbol{\beta}}_j\in\hat{\mathcal{V}}_k\right\},\ k\in\{1,\ldots,K_0\}. \tag{12}$$

To rewrite $\hat{G}_n$ in (10) as summation from 1 to $K_0$ instead of 1 to $K$, we need the following additional notations. Let

$$\hat{\alpha}(k_1,\ldots,k_T)=\sum_{S_1\in\hat{\mathcal{I}}_{k_1}}\cdots\sum_{S_T\in\hat{\mathcal{I}}_{k_T}}\hat{\pi}_{S_1}\hat{P}_{S_1S_2}\times\cdots\times\hat{P}_{S_{T-1}S_T}, \tag{13}$$

where $k_1,\ldots,k_T\in\{1,\ldots,K_0\}$ represent an estimate of the probability of a specific state trajectory $\{k_1,\ldots,k_T\}$. Similar to [15], we utilize $\hat{\alpha}_m$ to estimate $\pi_m$ by

$$\hat{\alpha}_k=\sum_{k_2=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\hat{\alpha}(k_1=k,k_2,\ldots,k_T),\ k\in\{1,\ldots,K_0\}. \tag{14}$$

We also define

$$\hat{H}\left(k_1,\ldots,k_T,\boldsymbol{\beta}^S\right)=\frac{\sum_{S_1\in\hat{\mathcal{I}}_{k_1}}\cdots\sum_{S_T\in\hat{\mathcal{I}}_{k_T}}\hat{\pi}_{S_1}\hat{P}_{S_1S_2}\times\cdots\times\hat{P}_{S_{T-1}S_T}\delta(\hat{\boldsymbol{\beta}}^S)}{\hat{\alpha}(k_1,\ldots,k_T)},$$

where $k_1,k_2,\ldots,k_T\in\{1,\ldots,K_0\}$, as the estimate of a mixing measure in its own right. Similar to (14), we define

$$\hat{H}_k=\sum_{k_2=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\hat{H}\left(k_1=k,k_2,\ldots,k_T,\boldsymbol{\beta}^S\right),\ k\in\{1,\ldots,K_0\}. \tag{15}$$

With the preceding notations, $\hat{G}_n(\boldsymbol{\beta}^S)$ in (10) can be rewritten as

$$\hat{G}_n(\boldsymbol{\beta}^S) = \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \hat{\alpha}\,(k_1, \ldots, k_T)\, \hat{H}\left(k_1, \ldots, k_T, \boldsymbol{\beta}^S\right), \tag{16}$$

which assembles different estimates $\hat{\boldsymbol{\beta}}_j$ in $\hat{\mathcal{V}}_k$ to form the state trajectories that only have $K_0$ states.

Under the representation in (16), we derive the theoretical properties of $\hat{G}_n(\boldsymbol{\beta}^S)$, where the $\boldsymbol{\beta}^S$ in parenthesis may be omitted if the context is clear. Denote by $G_0$ the true mixing measure. Theorem 1 shows that $\hat{G}_n$ is an consistent estimator of $G_0$ and $\left\{\hat{\mathcal{V}}_k\right\}$, $k \in \{1, \ldots, K_0\}$, forms a cluster partition of $\left\{\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K\right\}$. Theorem 2 establishes the consistency of order selection of RHMM.

**Theorem 1.** *Suppose that RHMM is identifiable and $F(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi})$ satisfies the mild regular conditions stated in Appendix A. If $\lambda_n = cn^{-\frac{1}{4}} \log n$ for SCAD penalty and some $c > 0$. Then, we have the following:*

*(1) For any continuous point of $\boldsymbol{\beta}^S$ of $G_0$, we have $\hat{G}_n(\boldsymbol{\beta}^S) \xrightarrow{p} G_0(\boldsymbol{\beta}^S)$.*

*(2) $\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$ and $\hat{\alpha}_k = \pi_{0k} + o_p(1)$ for all $k \in \{1, \ldots, K_0\}$. Furthermore, for each $\ell \in \{1, \ldots, K\}$, a unique $k \in \{1, \ldots, K_0\}$ exists, such that $\|\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_{0k}\|_2 = o_p(1)$. Thus, $\left\{\hat{\mathcal{V}}_k\right\}$, $k \in \{1, \ldots, K_0\}$, is a cluster partition of $\left\{\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K\right\}$ in probability.*

Notably, even though we already have the consistency of $\hat{G}_n$ in Theorem 1, it does not necessarily lead to the consistency of order selection. With an additional condition in Theorem 2, we can show the probability that each $\hat{\mathcal{V}}_k$ and $\hat{H}_k$ for all $k = 1, 2, \ldots, K_0$ have a unique element tends to one, which is also known as order selection consistency.

**Theorem 2.** *We assume that the same conditions in Theorem 1 hold. Under the true dynamic finite mixture density $F(\boldsymbol{Y}; G_0)$, if $\hat{G}_n$ falls into an $O(n^{-\frac{1}{4}})$ neighborhood of $G_0$, then $P\left(\hat{K}_n = K_0\right) \to 1$ as $n \to \infty$.*

The proofs of the above theorems are provided in Appendix B.

**Remark 3.** The assumption that $\hat{G}_n$ falls into an $O(n^{-\frac{1}{4}})$ neighborhood of $G_0$ is connected to the well-known results by [7] on finite mixture models. Chen [7] showed that the optimal rate of estimating the finite mixing distribution $G_0$ is $O(n^{-1/4})$ when the order is unknown but becomes $O(n^{-1/2})$ when the order is known. Considering that the proposed RHMM can be reformulated as a finite mixture model with an unknown order, the same rate $O(n^{-1/4})$ is expected. Moreover, the non-convex SCAD penalty is proven to achieve the oracle property [12]. Such an appealing feature of the SCAD penalty may even enable the proposed MPLE to enjoy the same convergence rate $O(n^{-1/2})$ as if the order is known a priori. Thus, assuming the rate of $O(n^{-1/4})$ in our study seems justifiable. Our simulation results in Table S3 (Online Supplementary Material S2) further confirm the feasibility of this assumption.

## 3. Estimation

In the presence of hidden states, the expectation–maximization (EM) algorithm together with the Baum–Welch algorithm [4] is known as an efficient statistical estimation method to obtain the maximum likelihood estimate of $\boldsymbol{\Psi}$. In this section, we propose an ECM–ITD algorithm to obtain the MPLE $\hat{\boldsymbol{\Psi}}_n$ of $\boldsymbol{\Psi}$ in RHMM.

### *3.1. ECM–ITD Algorithm*

Under the EM algorithm terminology, with the help of forward-backward technique, we propose the ECM updating steps as follows:

$$\pi_k^{(p+1)} = \frac{\sum_{i=1}^{n} h^{(p)}(S_{i1} = k) + C_k}{n + KC_k},\ P_{rs}^{(p+1)} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} h^{(p)}(S_{i,t-1} = r, S_{it} = s)}{\sum_{i=1}^{n} \sum_{t=2}^{T} h^{(p)}(S_{i,t-1} = r)}.$$

$$\boldsymbol{\beta}^{(p+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{k=1}^{K} \left[ \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(y_{it} | S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}) h^{(p+1)}(S_{it} = k) \right] - n \sum_{k=2}^{K} p_{\lambda_n}\left(\|\boldsymbol{\eta}_k\|_2\right),$$

where $k, r, s \in \{1, \ldots, K\}$ and $\boldsymbol{\eta_k} = \boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}$. The derivation can be found in Appendix C.

The maximization in updating $\boldsymbol{\beta}^{(p+1)}$ is an optimization problem with an objective function that includes the weighted log-likelihood and a nonsmooth and nonconvex SCAD penalty, which pose major challenges in optimization. Fan and Li [12] recommended a locally quadratic approximation (LQA) for SCAD penalty to make the optimization feasible. However, LQA-based methods cannot provide precisely sparse solutions due to a ridge-type penalized solution in each iteration. Zou and Li [45] proposed a well-performed local linear approximation (LLA) method for SCAD. Hung et al. [15] adopted this method for their Gaussian HMMs to deal with a similar optimization problem.

She [35] developed an iterative threshold algorithm for penalized generalized linear models with grouped predictors, in which the optimization problem is similar to that in updating $\boldsymbol{\beta}^{(p+1)}$. Xu and Chen [40] further proposed an ITD algorithm for order selection in finite mixture models, but they restricted a special case where the mixture components were determined by a one-dimensional parameter. The basic idea of the preceding thesholding-based methods is to decompose the original optimization problem into a set of easy problems, for which the solutions can be obtained by a soft-threshold operation, thereby leading to a sparse estimate directly and achieving the aim of order selection faithfully.

We develop an extended ITD algorithm in our multidimensional setting. First, we impose a constraint $\boldsymbol{\eta}_1 = \boldsymbol{\beta}_1$ to form a one-to-one mapping between $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$ and $\boldsymbol{\beta}$, i.e., $\boldsymbol{\beta}_k = \sum_{l=1}^{k} \boldsymbol{\eta}_l$, $k \in \{1, \ldots, K\}$. Using model (1), we convert the optimization of updating $\boldsymbol{\beta}^{(p+1)}$ as

$$\boldsymbol{\eta}^{(p+1)} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left\{ G(\boldsymbol{\eta}) = -\sum_{k=1}^{K} \varphi_k\Big(\sum_{l=1}^{k} \boldsymbol{\eta}_l\Big) + n \sum_{k=2}^{K} p_{\lambda_n}(\|\boldsymbol{\eta}_k\|_2) \right\}, \tag{17}$$

where $\varphi_k(\boldsymbol{\beta}_k) = \sum_{i=1}^{n} \sum_{t=1}^{T} \{y_{it}\theta_{itk} - b(\theta_{itk})\} h^{(p+1)}(S_{it} = k)$ and $\theta_{itk} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_k$. In the following, we denote $h^{(p+1)}(S_{it} = k)$ as $h_{itk}$ by suppressing the superscript to simplify notation.

Let $\boldsymbol{\zeta}_k = \sum_{l=1}^{k} \boldsymbol{\xi}_l$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)$, $k \in \{1, \ldots, K\}$. Inspired by the prevailing iterative shrinkage-thresholding algorithm (ISTA) for regulated convex optimization problem, we optimize a surrogate function $\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m)})$, which shares a similar ascent property in minorization–maximization (MM) algorithm, to downhill the value of $G(\boldsymbol{\eta})$ iteratively. The surrogate function $\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta})$ is defined as follows:

$$\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}) = \rho G(\boldsymbol{\xi}) + \frac{1}{2} \sum_{j=1}^{K} \|\boldsymbol{\xi}_j - \boldsymbol{\eta}_j\|_2^2 - \rho \left[ \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \Big\{ b(\boldsymbol{x}_{it}^T \boldsymbol{\zeta}_k) - b(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k) - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k) \Big[ \boldsymbol{x}_{it}^T (\boldsymbol{\zeta}_k - \boldsymbol{\beta}_k) \Big] \Big\} h_{itk} \right], \tag{18}$$

where $\rho$ is a positive scale parameter, and $b(\cdot)$ and $b'(\cdot)$ are the known functions defined in (1). The construction of (18) is similar to the formula in ISTA, but the third term in the right-hand side of (18) differs from that in ISTA. The extended ITD algorithm updates $\boldsymbol{\eta}$ through $\boldsymbol{\eta}^{(m+1)} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\ \tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m)})$, until the algorithm converges.

Notably, $\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta})$ can be further decomposed into additive components of $\boldsymbol{\xi}_j$, $j \in \{1, \ldots, K\}$, if we expand $G(\boldsymbol{\xi})$. This additive property facilitates the decomposition of the updating process into $K$ simple sub-problems: $j \in \{2, \ldots, K\}$,

$$\begin{aligned} \boldsymbol{\eta}_1^{(m+1)} &= \underset{\boldsymbol{\xi}_1}{\operatorname{argmin}} \left\| \boldsymbol{\xi}_1 - \left[ \boldsymbol{\eta}_1^{(\boldsymbol{m})} + \rho \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \Big( y_{it} - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k^{(m)}) \Big) \boldsymbol{x}_{it} h_{itk} \right] \right\|_2^2 \\ \boldsymbol{\eta}_j^{(m+1)} &= \underset{\boldsymbol{\xi}_j}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{\xi}_j - \left[ \boldsymbol{\eta}_j^{(\boldsymbol{m})} + \rho \sum_{k=j}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \Big( y_{it} - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k^{(m)}) \Big) \boldsymbol{x}_{it} h_{itk} \right] \right\|_2^2 + \rho n p_{\lambda_n}\left(\|\boldsymbol{\xi}_j\|_2\right). \end{aligned} \tag{19}$$

By using Lemma 1 in [35], we can obtain optimal solution to (19) through a multivariate thresholding operator, which generalizes its univariate version proposed by [40]. Specifically, the optimal solution to the unified optimization problem $\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma} - \boldsymbol{z}\|_2^2 + \kappa p_{\lambda_n}(\|\boldsymbol{\gamma}\|_2)$ can be denoted as $\vec{\boldsymbol{S}}(\boldsymbol{z}; \kappa, a, \lambda_n)$ (Appendix D).

Thus, the optimizations in (19) can be implemented using multivariate thresholding operator $\vec{\boldsymbol{S}}(\cdot; 2n\rho, a, \lambda_n)$ as

follows:

$$\begin{aligned}\boldsymbol{\eta}_1^{(m+1)} &= \boldsymbol{\eta}_1^{(m)} + \rho \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ y_{it} - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k^{(m)}) \right\} \boldsymbol{x}_{it} h_{itk}, \\ \boldsymbol{\eta}_j^{(m+1)} &= \vec{\boldsymbol{S}}\Big[ \boldsymbol{\eta}_j^{(m)} + \rho \sum_{k=j}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ y_{it} - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k^{(m)}) \right\} \boldsymbol{x}_{it} h_{itk}; 2n\rho, a, \lambda_n \Big],\end{aligned} \tag{20}$$

$j \in \{2, \ldots, K\}$, and $\boldsymbol{\eta}^{(m)}$ is updated until the algorithm converges.

We emphasize that $\rho$ is a constant imposed to ensure the reduction of the value of $G(\boldsymbol{\eta})$ when we apply the updating procedure (20). The following theorem, which is similar to Theorem 3.1 in [40], formally states the conditions that ensure the convergence of the sequences of $\boldsymbol{\eta}^{(m)}$ and $\boldsymbol{\beta}^{(m)}$.

**Theorem 3.** *Assume that sequence $\boldsymbol{\eta}^{(m)}$ is generated from (20) and $\boldsymbol{\beta}_k^{(m)} = \sum_{l=1}^{k} \boldsymbol{\eta}_l^{(m)}$. Let $\tau_1$ be the maximum eigenvalue of $\sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{x}_{it} \boldsymbol{x}_{it}^T$ and $\tau_2^{(m)}$ be assigned to*

$$\tau_2^{(m)} = \max_{i,t,k} \sup_{0<\alpha<1} b'' \left\{ \boldsymbol{x}_{it}^T (\alpha \boldsymbol{\beta}^{(m+1)} + (1-\alpha) \boldsymbol{\beta}^{(m)}) \right\}. \tag{21}$$

*If $\rho^{-1} \geq K \tau_2^{(m)} \tau_1$, then $G\left(\boldsymbol{\eta}^{(m+1)}\right) \leq G(\boldsymbol{\eta}^{(m)})$. Furthermore, if space $\{\boldsymbol{\eta} : G(\boldsymbol{\eta}) \leq G\left(\boldsymbol{\eta}^{(0)}\right)\}$ is compact, then sequences $\{\boldsymbol{\eta}^{(m)}\}$ and $\{\boldsymbol{\beta}^{(m)}\}$ converge to a stationary point of $G(\boldsymbol{\eta})$.*

Theorem 3 ensures that the extended ITD algorithm carries out (17) efficiently at each iteration and finally leads to converged solutions in the inner iterations. Compared with the preceding alternative optimization methods, the extended ITD algorithm can intrinsically produce sparse solutions by using thresholding rules. Using the one-to-one map between $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, we can transfer $\boldsymbol{\eta}^{(m+1)}$ to $\boldsymbol{\beta}^{(m+1)}$ and then continue the outer iterations until convergence.

### *3.2. Some Practical Computational Issues*

We propose the following procedure to obtain the initial value of $\boldsymbol{\beta}$. First, we conduct a pilot study to obtain the MLE of $\boldsymbol{\beta}$ under (1) with a single hidden state, denoted as $\hat{\boldsymbol{\beta}}^{(0)}$. Then, we set $\boldsymbol{\beta}_k^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} (1 + c\boldsymbol{\epsilon}_k)$, $k \in \{1, \ldots, K\}$, where $\boldsymbol{\epsilon}_k \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_K)$, $c$ is a small constant to determine the magnitude of disturbance. For instance, $c = 0.5$ is used in the numerical analyses of this study. Then, we take the cluster ordering procedure in definition 3 to sort $\boldsymbol{\beta}_k^{(1)}$ as the formal initial value of $\boldsymbol{\beta}$ in the ECM–ITD procedure. In addition, the SCAD penalty is a non-convex penalty, leading to a non-convex objective function with multiple local solutions. Thus, multiple initial values are recommended to achieve a global solution.

As shown by Theorem 3, in the ITD algorithm, $\rho$ satisfying some specific condition ensures the convergence of the sequence of $G(\boldsymbol{\eta}^{(m)})$. Considering that the condition of $\rho$ should be satisfied in each iteration, we may set the value of $\rho$ as small as possible such that the condition can always hold. However, by checking the proof of Theorem 3, we know that $\rho$ controls the number of iterations required to converge; the larger $\rho$ we take, the faster the algorithm converges. In this sense, we should set $\rho$ as large as possible while maintaining the condition of $\rho$ at each iteration. To accelerate the extended ITD algorithm, we adapt the line-back search method to determine the value of $\rho$. Heuristically, a large value of $\rho_0$ is provided, which satisfies the condition initially. Then, at each inner iteration, namely, the $m$th iteration, we set $\rho^{(m)} = \rho_0$, and then multiply a constant $0 < \upsilon < 1$, which is often chosen to be 0.5 for convenience, to obtain $\rho^{(m)} = \upsilon \rho^{(m)}$ recursively until $G\left(\boldsymbol{\eta}^{(m+1)}\right) \leq G(\boldsymbol{\eta}^{(m)})$ holds.

Moreover, $y_{it}$ is frequently assumed to follow the normal distribution with the identity link function. In such circumstances, we can adapt a relaxation and asynchronous updating technique designed by [35] for Gaussian linear models with grouped predictors as follow:

$$\begin{aligned}&\boldsymbol{\xi}_\ell^{(m+1)} = (1-\omega) \boldsymbol{\eta}_\ell^{(m)} + \omega \left\{ \boldsymbol{\eta}_\ell^{(m)} + \rho \sum_{k=j}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} h_{itk} \left( y_{it} - b'(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k^{(m)}) \right) \boldsymbol{x}_{it} \right\}, \\ &\boldsymbol{\eta}_1^{(m+1)} = \boldsymbol{\xi}_1^{(m+1)} \text{ and } \boldsymbol{\eta}_j^{(m+1)} = \vec{\boldsymbol{S}}\left( \boldsymbol{\xi}_j^{(m+1)}; 2n\rho, a, \lambda_n \right),\end{aligned} \tag{22}$$

where $\ell \in \{1, \cdots, K\}$, and $j \in \{2, \cdots, K\}$.

As recommended by [35], $\omega = 2$ leads to a 40% reduction of the number of iterations in the experiment. Thus, we set this value as default in the analysis of Gaussian GLM. When $y_{it}$ follows non-Gaussian distributions, the original synchronous updating (19) is more efficient. Notably, the relaxation updating (22) degenerates to synchronous updating (19) when $\omega = 1$. Thus, we can use the relaxation updating with different $\omega$ to update $\boldsymbol{\eta}^{(m+1)}$ whichever distribution $y_{it}$ is sampled from.

The choice of upper bound $K$ is also crucial. A large value of $K$ causes high computational burden. Hence, if we have some prior knowledge about $K_0$ from the subject knowledge or existing literature, we can set $K$ to be slightly larger than $K_0$ as the upper bound to ensure $K > K_0$. In most substantive studies, an integer ranging from 5 to 10 may be enough for $K$. If $K < K_0$ occurs, the two penalties no longer work due to the nonexistence of redundant signals, leading to $\hat{K}_n = K$. Then, one can increase $K$ sequentially until $\hat{K}_n$ does not change anymore.

We use pseudo-code table (Online Supplementary Material S1) above to show the estimation procedure.

### *3.3. Choice of Tuning Parameters*

Tuning parameters influence the performance of order selection in the proposed RHMM. Although Theorems 1 and 2 guarantee that $\lambda_n = cn^{-\frac{1}{4}} \log n$ leads to the order selection consistency asymptotically, it does not provide a specific choice of the tuning parameter in practice. Hung et al. [15] used a half-sampling cross-validation (CV) method to selection $\lambda_n$ in the context of Gaussian HMM. However, the CV-type methods are time-consuming and impose a large burden in computation. A commonly used alternative method is the BIC-type criterion, which avoids computational intensity and offers the variable selection consistency [e.g., 39]. Given these desirable properties, we use the BIC-type criterion to select $\lambda_n$ in our proposed RHMM as follows:

$$\lambda_n = \underset{[\lambda_{\min}, \lambda_{\max}]}{\operatorname{argmin}} \; -2\ell_n(\hat{\boldsymbol{\Psi}}_n; \boldsymbol{Y}, \boldsymbol{X}) + |\hat{\boldsymbol{\Psi}}_n| \log n, \tag{23}$$

where $|\hat{\boldsymbol{\Psi}}_n|$ represents the number of parameters in $\hat{\boldsymbol{\Psi}}_n$. Empirically, we usually select $\lambda_{\min} = 0$ and $\lambda_{\max} = n^{-\frac{1}{4}} \log n$. This BIC-type criterion is well-performed, and the corresponding $\lambda_n$ eventually facilitates the excellent performance of the ECM-ITD procedure.

In addition to the choice of $\lambda_n$, two other tuning parameters, namely, $a$ in SCAD penalty and $C_K$ in (4), are involved in the proposed algorithm. Previous studies have reported that the choice of $a$ and $C_K$ are robust to the performance of statistical inference. Thus, we follow the literature to set these two tuning parameters. As recommended by [12] and [15], we set $a = 3.7$ and $C_K = 0.6 \log 10$ throughout the numerical studies in sections 4 and 5.

## 4. Simulation Study

### *4.1. Simulation 1*

We first consider RHMMs with $K_0 \in \{2, 3, 4\}$. For each setting of $K_0$, $y_{it}$ in state $k$ is generated from two cases: (1) a normal distribution with mean $\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k$ and standard deviation $\sigma_k = 0.25$ and (2) Poisson distribution with mean $e^{\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k}$. Covariates $\boldsymbol{x}_{it} = (x_{it1}, x_{it2}, x_{it3})$, where $x_{it1} = 1$, and $x_{it2}$ and $x_{it3}$ are independently generated from $N(0, 1)$ and $U(0, 1)$, respectively, where $U(0, 1)$ stands for the uniform distribution on (0,1). Two sample sizes, $n = 50$ and 100 for normal and $n = 100$ and 200 for Poisson, and a transition matrix with elements $P_{rs} = 1/K_0,\ r, s \in \{1, \ldots, K_0\}$ are considered. The state-specific regression coefficients for case (1) are set as follows: when $K_0 = 2$, $T = 4$, $\boldsymbol{\beta}_1 = (0, -0.5, 0.2)^T$, and $\boldsymbol{\beta}_2 = (0.5, 0, -0.2)^T$; when $K_0 = 3$, $T = 4$, $\boldsymbol{\beta}_1 = (0, 0.5, 0.2)^T$, $\boldsymbol{\beta}_2 = (0.5, 0.5, -0.2)^T$ and $\boldsymbol{\beta}_3 = (1, -0.5, 0.2)^T$; when $K_0 = 4$, $T = 6$, $\boldsymbol{\beta}_1 = (0, 1, 1.25)^T$, $\boldsymbol{\beta}_2 = (1, 2, 1)^T$, $\boldsymbol{\beta}_3 = (1.5, 1.25, 0.75)^T$, and $\boldsymbol{\beta}_4 = (2, 1, 1.5)^T$.

State-specific regression coefficients for Case (2) are similarly set as follows: when $K_0 = 2$, $T = 4$, $\boldsymbol{\beta}_1 = (0, 1, 0.5)^T$ and $\boldsymbol{\beta}_2 = (1, -0.5, 0)^T$; when $K_0 = 3$, $T = 4$, $\boldsymbol{\beta}_1 = (0, 0.5, 0.2)^T$, $\boldsymbol{\beta}_2 = (1, 1, 0.2)^T$, and $\boldsymbol{\beta}_3 = (2, 0.5, -0.2)^T$; when $K_0 = 4$, $T = 6$, $\boldsymbol{\beta}_1 = (-0.5, 0.3, 0.2)^T$, $\boldsymbol{\beta}_2 = (1.5, 0.5, 0.2)^T$, $\boldsymbol{\beta}_3 = (2.5, 0.3, 0.2)^T$, and $\boldsymbol{\beta}_4 = (3.3, 0.2, 0)^T$.

We adopt the BIC-type criterion (23) to select tuning parameter $\lambda_n$. Other less sensitive tuning parameters $a$ and $C_k$ are set to 3.7 and $0.6 \log 10$ as recommended. In addition, we take $\delta = 10^{-3}$, $c = 0.2$, and $\nu = 0.5$ in the ECM–ITD estimation procedure. For each simulation setting, we treat the increment of the log-likelihood less than $\delta$ and $\|\boldsymbol{\Psi}^{(p+1)} - \boldsymbol{\Psi}^{(p)}\|_2 < \delta$ simultaneously as convergence of the ECM–ITD algorithm. We initialize our ECM–ITD estimation procedure from a common upper bound $K = 7$ in all the scenarios considered. The larger $K_0$ is, the longer

**Table 1:** Proportions of order selection of the proposed ECM-ITD and existing methods for Case (1) in Simulation 1. Results are reported for two sample sizes, $n = 50$ and 100, on the basis of 500 replications

| $K_0$ | $\hat{K}_n$ | $n = 50$ | | | $n = 100$ | | |
|---|---|---|---|---|---|---|---|
| | | AIC | BIC | ECM–ITD | AIC | BIC | ECM–ITD |
| 2 | 2 | 0.488 | 0.998 | 1 | 0.636 | 1 | 1 |
| | 3 | 0.278 | 0 | 0 | 0.232 | 0 | 0 |
| | 4 | 0.234 | 0.002 | 0 | 0.132 | 0 | 0 |
| 3 | 2 | 0.010 | 0.578 | 0.344 | 0 | 0.072 | 0.026 |
| | 3 | 0.408 | 0.422 | 0.654 | 0.562 | 0.924 | 0.974 |
| | 4 | 0.376 | 0 | 0.002 | 0.304 | 0.004 | 0 |
| | 5 | 0.206 | 0 | 0 | 0.134 | 0 | 0 |
| 4 | 3 | 0.002 | 0.474 | 0.190 | 0 | 0.002 | 0.014 |
| | 4 | 0.614 | 0.524 | 0.808 | 0.690 | 0.980 | 0.984 |
| | 5 | 0.384 | 0.002 | 0.002 | 0.310 | 0 | 0.02 |

the ECM–ITD algorithm takes to converge. When $K_0 = 4$ and $n = 100$, the ECM–ITD in Case 1 takes less than 50 ECM steps to converge, and each ECM step takes around a second.

Tables 1 and Table S1 (Online Supplementary Material S2) show the empirical performance of order selection for the proposed method under Cases (1) and (2), respectively. Considering that only AIC and BIC are available for the order selection of RHMMs in the literature, we report the results obtained using AIC and BIC for comparison. The proportions of selecting different orders on the basis of 500 replications are reported in each table. The proposed method consistently outperforms AIC and BIC in all the simulation settings considered. The performance improves when the sample size increases but declines when true order $K_0$ increases. Moreover, the accuracy of detecting the correct order is slightly higher for normal responses than for Poisson responses.

### *4.2. Simulation 2*

This section examines whether different transition patterns affect the performance of proposed approach. To mimic the scenario of the ADNI study in real data analysis, we consider a larger RHMM with $K_0 = 5$, $T = 6$, six covariates, and two different transition matrices: (1) a general transition matrix $\boldsymbol{P}_1$ and (2) a band transition matrix $\boldsymbol{P}_2$ that only allows transitions between adjacent states. Here, $y_{it}$ in state $k$ is generated from a normal distribution with mean $\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k$ and standard deviation $\sigma_k = 0.25$, covariates $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{it6})^T$, where $x_{it1} = 1$, $x_{it2}$ are independently generated from $N(0, 1)$, and $x_{it3}$ to $x_{it6}$ are independently generated from $U(0, 1)$. Three sample sizes, $n = 100$, 200, and 400, are considered. The state-specific regression coefficients are assigned as $\boldsymbol{\beta}_1 = (0, 0, 0, 0, 0, 0)^T$, $\boldsymbol{\beta}_2 = (-1.5, 2.25, -1, 0, 0.5, 0.75)^T$, $\boldsymbol{\beta}_3 = (0.25, 1.5, 0.75, 0.25, -0.5, -1)^T$, $\boldsymbol{\beta}_4 = (-0.25, 0.5, -2.5, 1.25, 0.75, 1.5)^T$, and $\boldsymbol{\beta}_5 = (-1, -1.5, -0.25, 1.75, -0.5, 2)^T$. The settings of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ shown in Online Supplementary Material S2.

Likewise, the BIC-type criterion (23) is employed to select tuning parameter $\lambda_n$, and the convergence of the ECM–ITD algorithm is monitored by using $\delta = 10^{-3}$. We start the proposed estimation procedure from upper bound $K = 7$. Table S2 (Online Supplementary Material S2) presents the order selection results obtained using AIC, BIC, and the proposed ECM–ITD procedure on 200 replications.

Overall, the three methods perform not as well as in the previous settings. This result is expected because the RHMM is more complex due to the higher order and larger number of covariates in this simulation setting. Nevertheless, the proposed method selects the true order with fairly acceptable proportions and still considerably outperforms

AIC and BIC. When $n = 100$, the correct selection proportion of the proposed method is 20% higher than those of AIC and BIC, and such superiority becomes more pronounced with larger sample sizes. The order selection accuracy improves when sample size $n$ increases, which coincides with the selection consistency property in Theorem 2. All these results also indicate that the difficulty of detecting correct order increases dramatically when true order $K_0$ increases. Table S2 (Online Supplementary Material S2) further shows that the results obtained with band transition matrix $\boldsymbol{P}_2$ is slightly better than those obtained in general transition matrix $\boldsymbol{P}_1$.

Tables S3 and S4 (Online Supplementary Material S2) report the estimation results of the regression parameters obtained using the proposed and two-stage methods, respectively, under $K_0 = 5$ and general transition matrix $\boldsymbol{P}_1$. The two-stage approach selects the order through AIC or BIC in the first stage and performs parameter estimation with a fixed order in the second stage; suppose AIC or BIC correctly chooses the order. The two methods perform similarly, indicating that the bias brought by the proposed penalty method is negligible.

## 5. Real Data Analysis

We applied the proposed method to the dataset extracted from the ADNI study to demonstrate the empirical utility of our proposed method. The main goal is to detect the number of hidden phases of the neurodegenerative pathology. The ADNI was launched in 2003 and it collected serial magnetic resonance imaging, positron emission tomography, biospecimen biomarkers, and various clinical and neurocognitive measures from subjects under cognitive normal (CN) controls and subjects with mild cognitive impairment (MCI) or AD. For up-to-date and more detailed information, please refer to the official website `www.adni-info.org`.

We focused on 616 subjects collected from the ADNI-I, ADNI-II, and ADNI-Go study with four follow-up visits at baseline, 6 months, 12 months, and 24 months. Alzheimer's Disease Assessment Scale (ADAS) was devised to evaluate cognitive impairment in the assessment of AD [24], and ADAS-Cognitive 13 (ADAS13) consists of 13 items corresponding to words, spoken language, and simple commands. In this study, we treated ADAS13 as response $y_{it}$ and some clinical and generic variables as covariate $\boldsymbol{x}_{it}$ in the proposed RHMM. Covariate $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{it6})^T$, where $x_{it1} = 1$, $x_{it2}$: age at each visit, $x_{it3}$: gender (1 = female), $x_{it4}$: logarithm of the ratio of hippocampal volume over the whole brain volume (HIP), and $\{x_{it5}, x_{it6}\}$: apolipoprotein E (APOE)-$\varepsilon$4, which was coded as 0, 1, and 2, denoting the number of APOE-$\varepsilon$4 alleles.

We assumed that $y_{it}$ in given state $k$ followed a normal distribution with mean $\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_k$ and an unknown standard deviation $\sigma_k$. The continuous variables including $y_{it}$, $x_{i2}$, and $x_{it4}$ were standardized prior to analysis. Published medical reports [19, 21, 30] investigated the progression from CN to AD in increasingly refined ways. For example, an early studies [21] classified AD progression into three stages: CN, MCI, and AD. Later on, a more refined classification of MCI, namely, early MCI (EMCI) and late MCI (LMCI), becomes prevalent [19]. More recent study [30] further indicated an intermediate state between EMCI and CN, and the addition of this intermediate state to the study would be helpful in the refinement of patient care. Based on the existing literature, we expected that $K$ should not be too large, and thus we set a $K$ slightly larger than 5, say, $K = 7$, as the upper bound. Tuning parameter $\lambda_n$ was selected through BIC-type criterion (23), and two other tuning parameters $a$ and $C_k$ were assigned to the default value 3.7 and $0.6 \log 10$. Similarly, $c = 0.2$ and $\nu = 0.5$ were set in the ECM–ITD procedure, and $\delta = 10^{-3}$ was adopted to monitor the convergence of the algorithm. Using a 2.60-GHz Intel processor in MacBook Pro, the proposed procedure took 28.75 minutes to converge within 119 ECM steps with the tuning process.

Fig. S1 (Online Supplementary Matieral S3) provides the details of determining the tuning parameter $\lambda_n$ through BIC, where the left panel shows the plot of BIC values versus tuning parameter $\lambda_n$, and the right panel illustrates the estimated hidden states corresponding to tuning parameter $\lambda_n$. The minimum value of BIC was attained at $\lambda_n = 0.08$, and the corresponding estimated order $\hat{K}_n = 5$ was then selected. For comparison, AIC and BIC selected $\hat{K}_n = 6$ and 5, respectively. As shown in Theorem 2, the proposed ECM-ITD procedure is theoretically guaranteed. Table 3 presents the estimated state-specific coefficients along with their variability estimates obtained using parametric bootstrap. The estimated transition matrix are presented as Online Supplementary Material S3.

In addition, we used the well-known pseudo-residuals (see Chapter 6 of [46]) to conduct model checking for the proposed model. The pseudo residual $z_{it}$ is defined as follows:

$$z_{it} = \Phi^{-1}\left[ \sum_{k=1}^{\hat{K}_n=5} h\left(S_{it} = k\right) \Phi\Big(\frac{y_{it} - \boldsymbol{x}_{it}^T \boldsymbol{\beta}_{S_{it}}}{\sigma_{S_{it}}}\Big)\right], \quad i \in \{1, \ldots, n\},\ t \in \{1, \ldots, T\},$$

**Table 2:** Five estimated hidden states, and the state-specific regression coefficients with standard errors in parentheses in the conditional regression model for the ADNI study

| | State | | | | |
|---|---|---|---|---|---|
| Par. | 1 (CN) | 2 (SMC) | 3 (EMCI) | 4 (LMCI) | 5(AD) |
| $\beta_{k1}$ | -1.007(0.023) | -0.628(0.031) | -0.033(0.029) | 0.876(0.035) | 2.225(0.167) |
| $\beta_{k2}$ | 0.092(0.017) | 0.096(0.019) | 0.069(0.023) | 0.076(0.027) | 0.034(0.068) |
| $\beta_{k3}$ | 0.037(0.022) | 0.111(0.026) | 0.023(0.036) | -0.066(0.052) | -0.464(0.168) |
| $\beta_{k4}$ | -0.146(0.013) | -0.257(0.021) | -0.378(0.017) | -0.430(0.025) | -0.347(0.105) |
| $\beta_{k5}$ | 0.057(0.036) | 0.203(0.042) | 0.192(0.037) | 0.089(0.046) | 0.519(0.215) |
| $\beta_{k6}$ | 0.278(0.214) | 0.445(0.236) | 0.407(0.145) | 0.184(0.214) | 0.240(0.409) |
| $\sigma_k$ | 0.227(0.010) | 0.266(0.009) | 0.315(0.011) | 0.377(0.015) | 0.770(0.049) |

where the item inside $\Phi^{-1}(\cdot)$ is the conditional cumulative fitted density function of $y_{it}$ given all except $y_{it}$, and $h(S_{it} = k)$ is the fitted version of (34). If the model is correctly specified, the pseudo-residuals $\{z_{it}\}, i \in \{1, \ldots, n\}$, $t \in \{1, \ldots, T\}$ should be realizations of the standard normal distribution. Fig. S2 in Online Supplementary Material S3 depicts the QQ plot for pseudo-residuals versus standard normal quantiles, suggesting that the proposed model is plausible.

We present the following observations. First, state-specific intercept $\beta_{k1}$ exhibits an ascending trend; patients had the lowest ADAS13 score in state 1, and the highest score in state 5. As ADAS13 measures cognitive impairment with a high score indicating low cognitive ability [28], states 1 to 5 can be explained as CN, significant memory concern (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD accordingly. This classification has been reported in the public literature and ADNI study from ADNI-II to the latest phase [19]. Some existing studies identify four (CN, EMCI, LMCI, AD) instead of five states. The ADNI-II study suggests that SMC is highly relevant to the AD progression and introducing an additional SMC state minimizes the stratification of cognitive ability and fills the gap between CN and EMCI. Published reports argued that SMC could address the vague demarcation between CN and EMCI [30].

Second, the effects of age and gender on ADAS13 ($\beta_{k2}$ and $\beta_{k3}$) are almost all close to 0. The exception is $\beta_{53} = -0.4637(0.1684)$, which implies that in AD state women suffer from more serious cognitive impairment than men. The medical literature also reported this phenomenon [29]. The effect of HIP on ADAS13, $\beta_{k4}$, is negative in all the states, and exhibits a nearly descending trend. The negative effect indicates that a high hippocampus volume is associated with good memory. This result coincides with the common sense that the hippocampus is crucial in transferring the short-term memory to long-term memory. The nearly descending trend of the effects implies that the atrophy of hippocampal volume increasingly impairs cognitive ability in the progression of AD from early to late stage. This finding concurs with public reports that atrophy of hippocampus is significantly correlated with cognitive decline [10]. The effects of APOE-$\varepsilon$4 alleles, $\beta_{k5}$ and $\beta_{k6}$, are all positive across states, and the magnitude of $\beta_{k6}$ is larger than that of $\beta_{k5}$ in majority of the states. This finding implies that carrying APOE-$\varepsilon$4 alleles increases the AD risk, and carrying two alleles in general has a more pronounced effect compared with carrying only one allele. We also notice that carrying APOE-$\varepsilon$4 alleles plays the most prominent role in impairing cognitive function when patients are in the SMC or EMCI phase, which aligns with the medical report that APOE-$\varepsilon$4 alleles are a crucial biomarker in SMC and EMCI [17, 30, 31].

Lastly, the estimated transition matrix $\hat{\boldsymbol{P}}$ exhibits a band structure. Patients' cognitive impairment transits only between adjacent states in the pathology of AD. Patients stay in their current states or transit to a severe state with over 95% probability, indicating that the progression of AD is irreversible. $\hat{P}_{55} = 0.995$ further shows that AD is near to be absorbing state, which reconfirms the irreversibility of AD revealed by previous studies [14].

## 6. Discussion

In this study, we have proposed a novel likelihood-based double penalized method coupled with an efficient ECM–ITD algorithm to conduct order selection for RHMM. Moreover, we have established theoretical results to show that the proposed method enjoys order selection consistency under mild regularity conditions and to ensure convergence of the ECM–ITD algorithm. To the best of our knowledge, this study is the first to investigate the order selection problem and its asymptotic properties in the context of RHMMs.

The present work has limitations. First, we assume that the transition is homogeneous. This assumption may be restrictive in practice because between-state transitions are frequently influenced by certain covariates, thereby leading to heterogenous transitions. A continuation-ratio logit model [1, 36] can be considered to model heterogeneous transitions. However, whether the theoretical results developed in this study still hold in the presence of heterogeneous transition remains unknown and require further investigation. Second, the present study focuses on the order selection of RHMM. In many substantive studies, selection of potential predictors is also interesting, especially when the dimension of predictors is high. Thus, the proposed approach may be generalized to include additional penalties on regression parameters to simultaneously conduct order and variable selection. Third, the conditional regression model in the proposed RHMM only accommodates scalar predictors. Extending our method to incorporate functional or image predictors can considerably improve model flexibility and utility. Finally, developing a highly flexible RHMM with time-varying regression parameters is of considerable interest and worth considering in the future. These possible extensions raise new theoretical and computational challenges.

### Acknowledgments

This research was fully supported by GRF grants 14301918 and 14302519 from the Research Grant Council of the Hong Kong Special Administrative Region. We thank the Editor, the Associate Editor, and two anonymous reviewers for their helpful suggestions that greatly improve this paper.

### Appendix A: Identifibiality and Regularity Conditions for RHMM

The identifiability conditions for RHMM in Theorems 1 and 2 are extended from the previous work of [15] and [26], with the former focusing on original HMMs excluding the content of regression in the emission probability. The RHMM is strongly identifiable if the following conditions are satisfied:

$A_1$. The Markov chain $\{\boldsymbol{S_i}\}$ is ergodic, that is, irreducible and aperiodic.

$A_2$. The parameter space $\boldsymbol{\Omega}$ for state-specific coefficients $\boldsymbol{\beta}_{S_{it}}$ is compact.

$A_3$. $f(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}})$ is continuous in each element of $\boldsymbol{\beta}_{S_{it}}$.

$A_4$. Given $\epsilon > 0$, there exists $A > 0$ such that for every $\boldsymbol{\beta}_{S_{it}} \in \boldsymbol{\Omega}$, $f(A; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}}) - f(-A; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}}) \geq 1 - \epsilon$.

$A_5$. Given covariate $\boldsymbol{X}$, the family of finite mixture models with covariates of $\{f(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}})\}$ in RHMM is strongly identifiable, such that $\left\| \sum_{S_{it}} \left\{ \zeta_j f(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}}) + \boldsymbol{\xi}_{S_{it}}^{\top} \frac{\partial f(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}})}{\partial \boldsymbol{\beta}_{S_{it}}} + \boldsymbol{\gamma}_{S_{it}}^{\top} \frac{\partial^2 f(y_{it}; \boldsymbol{x}_{it}, \boldsymbol{\beta}_{S_{it}})}{\partial \boldsymbol{\beta}_{S_{it}} \partial \boldsymbol{\beta}_{S_{it}}^{\top}} \boldsymbol{\gamma}_{S_{it}} \right\} \right\|_{\infty} = 0$, then $\zeta_k = 0$ and $\boldsymbol{\xi}_k = \boldsymbol{\gamma}_k = \boldsymbol{0}$, $k \in \{1, \ldots, K\}$.

$A_6$. The upper bound real number of state $K$ exists and is finite.

The above identifibiality conditions are natural, and most members in the exponential family satisfy it. Taking exponential family density function with covariates as the emission part can easily satisfy these conditions.

Similarly, the regularity conditions for RHMM, which are also extended from [6, 8, 15], state that

$B_1$. (a) $E(|\log F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi})|) < \infty$.

(b) there exists $\omega > 0$ such that $F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}, \omega)$ is measurable and

$$E(|\log F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}, \omega)|) < \infty,$$

where $F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}, \omega) = 1 + \sup_{\|\boldsymbol{\Psi}' - \boldsymbol{\Psi}\|_2 \le \omega} F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}')$.

$B_2$. $F(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi})$ is differentiable with respect to $\boldsymbol{\Psi}$ to order 3. Moreover, the derivatives are continuous in $\boldsymbol{\Psi}$.

$B_3$. Let $U_{i,\beta^{n_1}_{k(1),l_1},\ldots,\beta^{n_t}_{k(t),l_t}}(\boldsymbol{\beta}, G) = \{\frac{\partial^{n_1+\cdots+n_T}}{\partial\beta^{n_1}_{k(1),l_1}\cdots\partial\beta^{n_t}_{k(t),l_t}} f(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\beta}^{\{\boldsymbol{k}_l\}})\}/F(\boldsymbol{Y}_i, G)$ where $n_1 + \cdots + n_T \le 3$. For each atom of $G_0, \boldsymbol{\beta}_0$, there will exists a small neighborhood of $\boldsymbol{\beta}_0$ and a function $q(\boldsymbol{Y})$ such that $E[q^2(\boldsymbol{Y})] < \infty$ and $G, G', \boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ in the neighborhood, we have

$$\left|U_{i,\beta^{n_1}_{k(1),l_1},\ldots,\beta^{n_t}_{k(t),l_t}}(\boldsymbol{\beta}, G) - U_{i,\beta^{n_1}_{k(1),l_1},\ldots,\beta^{n_t}_{k(t),l_t}}(\boldsymbol{\beta}', G')\right| \le q(\boldsymbol{Y}_i)\{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2 + |G - G'|\}.$$

$B_4$. The matrix with the $(k_1, k_2)$th element $E[U_{1,\beta_{m,l}}(\boldsymbol{\beta}_{0k_1}, G_0)U_{1,\beta_{m,l}}(\boldsymbol{\beta}_{0k_2}, G_0)]$ is finite and positive definite.

$B_5$. $\boldsymbol{\beta}_{0S_{it}}$ are the interior points in $\boldsymbol{\Omega}$.

Conditions $B_1$ to $B_5$ ensure that the ordinary maximum likelihood estimator of $\hat{G}_n$ with known $K_0$ is $\sqrt{n}$-consistent and asymptotically normal.

## Appendix B: Proof of Theorems

This section provides the technical details of proofs of Theorem 1, 2 and 3.

**Proof of Theorem 1.** First we prove part (1). By Jensen's inequality, for any $(G_0, \boldsymbol{\beta}_0) \neq (G, \boldsymbol{\beta})$,

$$\frac{1}{n}[\ell_n(G, \boldsymbol{\beta}) - \ell_n(G_0, \boldsymbol{\beta}_0)] \xrightarrow{a.s.} E\left[\log \frac{F(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi})}{F(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi}_0)}\right] < \log\left[E\frac{F(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi})}{F(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi}_0)}\right] = 0$$

holds under regularity condition $B_1$ and identifiability conditions. Thus, there exist a constant $C > 0$ such that $l_n(G, \boldsymbol{\beta}) - l_n(G_0, \boldsymbol{\beta}_0) \le -Cn$ for large $n$. Owing to the compact assumption, i.e., condition $A_2$ holds, there exist $N > 0$ finite open coverage of $\boldsymbol{\Omega}$ and for any compact N not containing $(G_0, \boldsymbol{\beta}_0)$,

$$\sup_N \ell_n(G, \boldsymbol{\beta}) - \ell_n(G_0, \boldsymbol{\beta}_0) \le -Cn.$$

Also by the property of SCAD penalty, for $\lambda_n = O(n^{-\frac{1}{4}} \log n)$ and $a > 2$, we have $p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_{\boldsymbol{k}}\right\|_2\right) = o(1)$. Thus,

$$\begin{aligned}
\sup_N \tilde{\ell}_n(G, \boldsymbol{\beta}) - \tilde{\ell}_n(G_0, \boldsymbol{\beta}_0) &\le \sup_N \Big[\ell_n(G, \boldsymbol{\beta}) - \ell_n(G_0, \boldsymbol{\beta}_0) + C_k \sum_{k=1}^{K} \log \pi_k + n \sum_{k=2}^{K_0} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_{\boldsymbol{0k}}\right\|_2\right)\Big] \\
&= \sup_N \left\{\ell_n(G, \boldsymbol{\beta})\Big[1 + \frac{\frac{1}{n}C_k \sum_{k=1}^{K} \log \pi_k}{\frac{1}{n}\ell_n(G, \boldsymbol{\beta})}\Big] - l_n(G_0, \boldsymbol{\beta}_0)\Big[1 + \frac{\sum_{k=2}^{K_0} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_{\boldsymbol{0k}}\right\|_2\right)}{\frac{1}{n}l_n(G_0, \boldsymbol{\beta}_0)}\Big]\right\} \\
&= \sup_N \ell_n(G, \boldsymbol{\beta})(1 + o_p(1)) - \ell_n(G_0, \boldsymbol{\beta}_0)(1 + o_p(1)) \le -Cn
\end{aligned}$$

holds for large n. Together with MPLE $(\hat{G}_n, \hat{\boldsymbol{\beta}}) = \text{Argmax}\, \tilde{\ell}_n(G, \boldsymbol{\beta})$, we have $\hat{G}_n \xrightarrow{p} G_0$, which has at least $K_0$ atom.

Next, we turn to part (2). Because the $\hat{G}_n$ is a consistent estimator of $G_0$, interpreting this consistency result accordingly, we get $\hat{\alpha}_k = \pi_{0k} + o_p(1)$ immediately. Also we assume that $(\bar{G}_n, \bar{\boldsymbol{\beta}})$ is the ordinal MLE of $\ell_n(G_n, \boldsymbol{\beta})$ . by

the optimality of $\ell_n(\hat{G}_n, \hat{\boldsymbol{\beta}})$ and similar argument above, we have:

$$
\begin{aligned}
0 \leq \tilde{\ell}_n(\hat{G}_n, \hat{\boldsymbol{\beta}}) - \tilde{\ell}_n(G_0, \boldsymbol{\beta}_0) &\leq \ell_n(\hat{G}_n, \hat{\boldsymbol{\beta}}) - \ell_n(G_0, \boldsymbol{\beta}_0) + C_k \sum_{k=1}^{K} \log \hat{\pi}_k - C_k \sum_{k=1}^{K} \log \pi_{0k} + n \sum_{k=2}^{K_0} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_{\boldsymbol{0k}}\right\|_2\right) \\
&\leq \ell_n(\bar{G}_n, \bar{\boldsymbol{\beta}}) - \ell_n(G_0, \boldsymbol{\beta}_0) + C_k \sum_{k=1}^{K} \log \hat{\pi}_k - C_k \sum_{k=1}^{K} \log \pi_{0k} + n \sum_{k=2}^{K_0} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_{\boldsymbol{0k}}\right\|_2\right) \\
&\leq \{\ell_n(\bar{G}_n, \bar{\boldsymbol{\beta}}) - \ell_n(G_0, \boldsymbol{\beta}_0)(1 + o_p(1))\} + C_k \sum_{k=1}^{K} \log \hat{\pi}_k - C_k \sum_{k=1}^{K_0} \log \pi_{0k} = O_p(1) + C_k \sum_{k=1}^{K} \log \hat{\pi}_k - C_k \sum_{k=1}^{K_0} \log \pi_{0k},
\end{aligned}
$$

where the last equality holds due to the strongly identifiability assumption A5 [7, 9]. Therefore, $C_k \sum_{k=1}^{K} \log \hat{\pi}_k \geq O_p(1) + C_k \sum_{k=1}^{K_0} \log \pi_{0k} = O_p(1)$, which implies the estimated $\hat{\pi}_k$ are all strictly positive in probability. Thus, by the definition of of $\hat{H}_k$, we conclude that the atoms of $\hat{H}_k$ converges in probability to $\boldsymbol{\beta}_{0k}$, $k \in \{1, \ldots, K_0\}$, i.e. for each $\ell \in \{1, \ldots, K\}$, there exists a unique $k \in \{1, \ldots, K_0\}$, such that $\|\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_{0k}\|_2 = o_p(1)$. □

Before formally performing the proof of Theorem 2, we introduce a useful result from [34].

**Lemma 1.** *Let $h(\boldsymbol{Y}; \boldsymbol{\beta})$ be continuous at $\boldsymbol{\beta}_{\boldsymbol{0}}$, uniformly in $\boldsymbol{Y}$. Let $F$ be a distribution function for which $\int |h(\boldsymbol{Y}; \boldsymbol{\beta})| dF(\boldsymbol{Y}) < \infty$. Let $\boldsymbol{\beta}_1^n = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$ be a random sample from $F$ and suppose that $T_n = T_n(\boldsymbol{\beta}_1^n)$ is a function of the sample such that $T_n \xrightarrow{p} \boldsymbol{\beta}_0$. Thus, we have*

$$
\frac{1}{n} \sum_{i=1}^{n} h\left(\boldsymbol{Y}_i; T_n\right) \xrightarrow{p} E_0 h\left(\boldsymbol{Y}; \boldsymbol{\beta}_0\right).
$$

**Proof of Theorem 2.** By theorem 1, it suffice to show that any estimate of G with $\hat{K}_n > K_0$ cannot be a local maximizer of $l_n(G, \boldsymbol{\beta})$, in an $n^{-\frac{1}{4}}$-neighborhood of $(G_0, \boldsymbol{\beta}_0)$. We suppose that $(\tilde{G}_n, \tilde{\boldsymbol{\beta}})$ is a maximizer of $\tilde{l}_n(G, \boldsymbol{\beta})$ among those with exact $K_0$ atoms and same mixing proportions $\alpha_1, \ldots, \alpha_{K_0}$. (Note that $\tilde{G}_n$ actually is constructed based on knowledge of $G_0$ instead of an estimator). Theorem 1 indicates that $\pi_k$ can be grouped and in each group they sum up to $\alpha_k$ and further $\alpha_k = \pi_{0k} + o_p(1)$ implying

$$
\sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K_0} \log \alpha_k < 0
$$

in probability. For the of SCAD penalty, as the similar arguments in [8, 15], we directly get:

$$
n \sum_{k=2}^{K} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_k\right\|_2\right) - n \sum_{k=2}^{K_0} p_{\lambda_n}\left(\left\|\tilde{\boldsymbol{\eta}}_{0k}\right\|_2\right) = n \sum_{k=1}^{K_0} \sum_{j, j+1 \in \mathcal{I}_k} p_{\lambda_n}\left(\left\|\boldsymbol{\eta}_j\right\|_2\right) \leq n \lambda_n \sum_{k=1}^{K_0} \sum_{j, j+1 \in \mathcal{I}_k} \left\|\boldsymbol{\eta}_j\right\|_2.
$$

This result indicates that there will exist extra penalty about the size of $n^{\frac{3}{4}} \log n$ times the discrepancies between the atoms in $\mathcal{I}_k$, compared to $\tilde{G}_n$, if some $\mathcal{I}_k$ contain more than one atom. Hence, the above two inequalities result in

$$
\tilde{\ell}_n(G, \boldsymbol{\beta}) - \tilde{\ell}_n(\tilde{G}_n, \boldsymbol{\beta}) \leq [\ell_n(G, \boldsymbol{\beta}) - \ell_n(\tilde{G}_n, \boldsymbol{\beta})] - n \lambda_n \sum_{k=1}^{K_0} \sum_{j, j+1 \in \mathcal{I}_k} \left\|\boldsymbol{\eta}_j\right\|_2. \tag{24}
$$

Now, we move to assess the order of difference of ordinal log-likelihood. We have

$$
\ell_n(G, \boldsymbol{\beta}) - \ell_n(\tilde{G}_n, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log(1 + \Delta_i) \leq \sum_{i=1}^{n} \Delta_i - \frac{1}{2} \sum_{i=1}^{n} \Delta_i^2 + \frac{1}{3} \sum_{i=1}^{n} \Delta_i^3, \tag{25}
$$

where,

$$\Delta_i = \frac{F\Big(\boldsymbol{Y}_i; G\Big) - F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)} = \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \alpha(k_1, \ldots, k_T) \frac{F\Big(\boldsymbol{Y}_i; H(k_1, \ldots, k_T, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}})\Big) - f\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)}$$
$$= \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \alpha(k_1, \ldots, k_T) \int \frac{f\Big(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}}\Big) - f\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)} dH(k_1, \ldots, k_T, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}})$$

and $\{\boldsymbol{k}_t\} = (k_1, \ldots, k_T)$. To circumvent the double index issue, in the following content, we take $k(1), \ldots, k(T)$ as the new notations for $k_1, \ldots, k_T$. By Taylor's Expansion, for any $\boldsymbol{\beta}^{\{\boldsymbol{k}_t\}}$ close enough to $\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}$ and some $\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}}$ between $\boldsymbol{\beta}^{\{\boldsymbol{k}_t\}}$ and $\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}$, we have

$$\frac{f\Big(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}}\Big) - f\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)} = \sum_{m=1}^{T} \sum_{\ell_1=1}^{q+1} (\beta_{k(m),\ell_1} - \tilde{\beta}_{k(m),\ell_1}) U_{i,\tilde{\beta}_{k(m),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n)$$
$$+ \frac{1}{2} \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m1,m2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} (\beta_{k(m1),\ell_1} - \tilde{\beta}_{k(m1),\ell_1})^{n_1} (\beta_{k(m2),\ell_2} - \tilde{\beta}_{k(m2),\ell_2})^{n_2} U_{i,\tilde{\beta}_{k(m1),\ell_1},\tilde{\beta}_{k(m2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n)$$
$$+ \frac{1}{6} \sum_{\substack{\ell_1,\ell_2,\ell_3 \in \{1,\ldots,q+1\} \\ m1,m2,m3 \in \{1,\ldots,T\} \\ n_1+n_2+n_3=3}} \left[ (\beta_{k(m1),\ell_1} - \tilde{\beta}_{k(m1),\ell_1})^{n_1} (\beta_{k(m2),\ell_2} - \tilde{\beta}_{k(m2),\ell_2})^{n_2} (\beta_{k(m3),\ell_3} - \tilde{\beta}_{k(m3),\ell_3})^{n_3} U_{i,\tilde{\beta}_{k(m1),\ell_1},\tilde{\beta}_{k(m2),\ell_2},\tilde{\beta}_{k(m3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n) \right].$$

For simplicity, we let

$$p(k(1)_{\ell_1}^{n_1} \ldots k(t)_{\ell_t}^{n_t}) = \int (\beta_{k(m1),\ell_1} - \tilde{\beta}_{k(m1),\ell_1})^{n_1} \times \cdots \times (\beta_{k(mt),\ell_t} - \tilde{\beta}_{k(mt),\ell_t})^{n_t} dH(k_1, \ldots, k_T, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}}).$$

Now, equipped with the above notations, we turns to bound the first term $\sum_{i=1}^{n} \Delta_i$,

$$\sum_{i=1}^{n} \Delta_i = \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \alpha(k_1, \ldots, k_T) \left[ \sum_{i=1}^{n} \frac{F\Big(\boldsymbol{Y}_i; H(k_1, \ldots, k_T, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}})\Big) - f\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)} \right],$$

Because $\tilde{G}_n$ is the maximizer of $\tilde{l}_n(G, \boldsymbol{\beta})$ with exact $K_0$ atoms, the partial derivatives of $\tilde{l}_n$ must varnish at $\tilde{G}_n$, and SCAD is flat at $\|\boldsymbol{\eta}\|_2 > 0$ for large n, it then follows that

$$\sum_{i=1}^{n} \frac{F\Big(\boldsymbol{Y}_i; H(k_1, \ldots, k_T, \boldsymbol{\beta}^{\{\boldsymbol{k}_t\}})\Big) - f\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}\right)}{F\left(\boldsymbol{Y}_i; \tilde{G}_n\right)} = \frac{1}{2} \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2}) \sum_{i=1}^{n} U_{i,\tilde{\beta}_{k(m_1),\ell_1},\tilde{\beta}_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n)$$
$$+ \frac{1}{6} \sum_{\substack{l_1,l_2,l_3 \in \{1,\ldots,q+1\} \\ m1,m2,m3 \in \{1,\ldots,T\} \\ n_1+n_2+n_3=3}} p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2} k(m_3)_{\ell_3}^{n_3}) \sum_{i=1}^{n} U_{i,\tilde{\beta}_{k(m_1),\ell_1},\tilde{\beta}_{k(m_2),\ell_2},\tilde{\beta}_{k(m_3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n).$$

Since $E\{\sum_{\substack{m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} U_{i,\tilde{\beta}_{k(m_1),\ell_1},\tilde{\beta}_{k(m_2),\ell_2}}(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}}, G_0)\} = 0$, by dominated convergence theorem, we have

$$\sum_{i=1}^{n} \sum_{\substack{m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} U_{i,\tilde{\beta}_{k(m1),l_1},\tilde{\beta}_{k(m2),l_2}}(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}}, G_0) = O_p(n^{\frac{1}{2}}), \quad \ell_1, \ell_2 \in \{1, \ldots, q+1\}.$$

Hence, we obtain

$$\sum_{i=1}^{n}\sum_{\substack{m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde{G}_n) \le nO_p(n^{-\frac{1}{4}}) + \sum_{i=1}^{n}\sum_{\substack{m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2}}(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}},G_0)$$
$$= O_p(n^{\frac{3}{4}}) + O_p(n^{\frac{1}{2}}) = O_p(n^{\frac{3}{4}})$$

under the condition $B_3$. Similarly, consider the third term, $\ell_1,\ell_2,\ell_3 \in \{1,\ldots,q+1\}$, we again have

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}} U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2},\tilde{\boldsymbol{\beta}}_{k(m_3),\ell_3}}(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}},G_0) = O_p(1).$$

Thus,

$$\frac{1}{6}\sum_{\substack{\ell_1,\ell_2,\ell_3\in\{1,\ldots,q+1\}\\ m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}k(m_3)_{\ell_3}^{n_3})\sum_{i=1}^{n} U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2},\tilde{\boldsymbol{\beta}}_{k(m_3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}},\tilde{G}_n) \tag{26}$$
$$= O_p(n)\sum_{\substack{\ell_1,\ell_2,\ell_3\in\{1,\ldots,q+1\}\\ m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}k(m_3)_{\ell_3}^{n_3}) = O_p(n)O_p(n^{-\frac{1}{4}})\sum_{\substack{m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2})$$
$$= O_p(n^{\frac{3}{4}})\sum_{\substack{m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}).$$

Combining these above results implies that there exist a constant $C_1$, such that

$$\sum_{i=1}^{n}\Delta_i \le C_1O_p(n^{\frac{3}{4}})\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\sum_{\substack{\ell_1,\ell_2\in\{1,\ldots,q+1\}\\ m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}}\alpha(k_1,\ldots,k_T)\,p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}). \tag{27}$$

Now, considering bounding the second term $\sum_{i=1}^{n}\Delta_i^2$, with similar argument in [27], we have

$$\sum_{i=1}^{n}\Delta_i^2 = \sum_{i=1}^{n}\Bigg\{\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Bigg[\sum_{\substack{\ell_1\in\{1,\ldots,q+1\}\\ m_1\in\{1,\ldots,T\}}} p(k(m_1)_{l_1})U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde{G}_n)$$
$$+\frac{1}{2}\sum_{\substack{\ell_1,\ell_2\in\{1,\ldots,q+1\}\\ m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2})U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde{G}_n)$$
$$+\frac{1}{6}\sum_{\substack{\ell_1,\ell_2,\ell_3\in\{1,\ldots,q+1\}\\ m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}k(m_3)_{\ell_3}^{n_3})U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2},\tilde{\boldsymbol{\beta}}_{k(m_3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}},\tilde{G}_n)\Bigg]\Bigg\}^2 = \text{(I)} + \text{(II)} + \text{(III)},$$

where

$$\text{(I)} = \sum_{i=1}^{n}\Bigg\{\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Bigg[\sum_{\substack{\ell_1\in\{1,\ldots,q+1\}\\ m_1\in\{1,\ldots,T\}}} p(k(m_1)_{\ell_1})U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde{G}_n)$$
$$+\frac{1}{2}\sum_{\substack{\ell_1,\ell_2\in\{1,\ldots,q+1\}\\ m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2})U_{i,\tilde{\boldsymbol{\beta}}_{k(m_1),\ell_1},\tilde{\boldsymbol{\beta}}_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde{G}_n)\Bigg]\Bigg\}^2,$$

$$\text{(II)} = \frac{1}{36}\sum_{i=1}^{n}\Bigg\{\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Bigg[\sum_{\substack{\ell_1,\ell_2,\ell_3\in\{1,\ldots,q+1\}\\ m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}}\Big(p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}k(m_3)_{\ell_3}^{n_3})U_{i,\tilde\beta_{k(m_1),\ell_1},\tilde\beta_{k(m_2),\ell_2},\tilde\beta_{k(m_3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Big)\Bigg]\Bigg\}^2,$$

$$\begin{aligned}\text{(III)} = \frac{1}{3}\sum_{i=1}^{n}\Bigg\{&\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Bigg[\sum_{\substack{\ell_1\in\{1,\ldots,q+1\}\\ m_1\in\{1,\ldots,T\}}}p(k(m_1)_{\ell_1})U_{i,\tilde\beta_{k(m_1),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\\ &+\frac{1}{2}\sum_{\substack{\ell_1,\ell_2\in\{1,\ldots,q+1\}\\ m_1,m_2\in\{1,\ldots,T\}\\ n_1+n_2=2}}p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2})U_{i,\tilde\beta_{k(m_1),\ell_1},\tilde\beta_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Bigg]\Bigg\}\\ &\times\Bigg\{\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Bigg[\sum_{\substack{\ell_1,\ell_2,\ell_3\in\{1,\ldots,q+1\}\\ m_1,m_2,m_3\in\{1,\ldots,T\}\\ n_1+n_2+n_3=3}}\Big(p(k(m_1)_{\ell_1}^{n_1}k(m_2)_{\ell_2}^{n_2}k(m_3)_{\ell_3}^{n_3})U_{i,\tilde\beta_{k(m_1),\ell_1},\tilde\beta_{k(m_2),\ell_2},\tilde\beta_{k(m_3),\ell_3}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Big)\Bigg]\Bigg\}.\end{aligned}$$

So that we denote $\boldsymbol{R}_{\{\boldsymbol{k}_t\},1} = \{p(k(m_1)_{\ell_1})\} \in \mathbb{R}^{T(q+1)\times 1}$, $\boldsymbol{R}_{\{\boldsymbol{k}_t\},2} = \{p(k(m_1)_{\ell_1}^{n_1}, k(m_2)_{\ell_2}^{n_2})\} \in \mathbb{R}^{3T^2(q+1)^2\times 1}$, where $n_1 + n_2 = 2$, $m_1, m_2 \in \{1,\ldots,T\}$ and $\ell_1,\ell_2 \in \{1,\ldots,q+1\}$. Similarly, we define $\boldsymbol{U}_{i,1}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n) = \{U_{i,\tilde\beta_{k(m_1),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\} \in \mathbb{R}^{T(q+1)\times 1}$, $\boldsymbol{U}_{i,2}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n) = \{U_{i,\tilde\beta_{k(m_1),\ell_1},\tilde\beta_{k(m_2),\ell_2}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\} \in \mathbb{R}^{3T^2(q+1)^2\times 1}$. Also we let

$$\boldsymbol{R}_j = (\boldsymbol{R}_{\{1,\ldots,1\},j},\ldots,\boldsymbol{R}_{\{K_0,\ldots,K_0\},j})^T,\quad \boldsymbol{V}_{ij}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n) = (\boldsymbol{U}_{i,j}(\tilde{\boldsymbol{\beta}}^{\{1,\ldots,1\}},\tilde G_n),\ldots,\boldsymbol{U}_{i,j}(\tilde{\boldsymbol{\beta}}^{\{K_0,\ldots,K_0\}},\tilde G_n))^T,\quad j\in\{1,2\}.$$

Then, $\boldsymbol{R} = (\boldsymbol{R}_1,\boldsymbol{R}_2)^T$, and $\boldsymbol{V}_i(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n) = (\boldsymbol{V}_{i1}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n),\boldsymbol{V}_{i2}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n))^T$.

Therefore, we could rewrite (I) as

$$\begin{aligned}\text{(I)} &= \sum_{i=1}^{n}\Bigg\{\sum_{k_1=1}^{K_0}\cdots\sum_{k_T=1}^{K_0}\alpha(k_1,\ldots,k_T)\Big[\boldsymbol{R}_{\{\boldsymbol{k}_t\},1}^T\boldsymbol{U}_{i,1}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)+\frac{1}{2}\boldsymbol{R}_{\{\boldsymbol{k}_t\},2}^T\boldsymbol{U}_{i,2}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Big]\Bigg\}^2\\ &\asymp\sum_{i=1}^{n}\Big\{\boldsymbol{R}_1^T\boldsymbol{V}_{i1}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)+\boldsymbol{R}_2^T\boldsymbol{V}_{i2}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Big\}^2\\ &=\sum_{i=1}^{n}\boldsymbol{R}^T\boldsymbol{V}_i(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\boldsymbol{V}_i^T(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\boldsymbol{R}=\boldsymbol{R}^T\Big\{\sum_{i=1}^{n}\boldsymbol{V}_i(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\boldsymbol{V}_i^T(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\Big\}\boldsymbol{R}.\end{aligned}$$

By lemma 1, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\boldsymbol{V}_i^T(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}},\tilde G_n)\xrightarrow{p}\boldsymbol{\Sigma}\triangleq E\{\boldsymbol{V}_1(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}},G_0)\boldsymbol{V}_1^T(\boldsymbol{\beta}_0^{\{\boldsymbol{k}_t\}},G_0)\}.$$

Hence,

$$\varrho_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{R}\|_F^2\lesssim\frac{1}{n}(I)\lesssim\varrho_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{R}\|_F^2$$

holds for large $n$ in probability. Furthermore, we have $\text{(I)} = O_p(n)\|\boldsymbol{R}\|_F^2$. Using the same argument, and noting that

$$(\beta_{k(m_1),\ell_1}-\tilde\beta_{k(m_1),\ell_1})^{n_1}(\beta_{k(m_2),\ell_2}-\tilde\beta_{k(m_2),\ell_2})(\beta_{k(m_3),\ell_3}-\tilde\beta_{k(m_3),\ell_3})=o_p(1)(\beta_{k(m_1),\ell_1}-\tilde\beta_{k(m_1),\ell_1})(\beta_{k(m_2),\ell_2}-\tilde\beta_{k(m_2),\ell_2})$$

for all $\boldsymbol{\beta}$ in $O(n^{-\frac{1}{4}})$ neighborhood of an atom of $G_0$, we get

$$\text{(II)} = o_p(n)\|\boldsymbol{R}\|_F^2.$$

Moreover, using Cauchy-Schwarz inequality, we have

$$|\text{(III)}|\le\sqrt{\text{(I)(II)}} = o_p(n)\|\boldsymbol{R}\|_F^2.$$

Therefore, these three order assessments conclude that there exist a constant $C_2 > 0$ such that

$$\sum_{i=1}^{n} \Delta_i^2 \geq C_2 n \|\boldsymbol{R}\|_F^2 = C_2 n \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \Bigg[ \sum_{\substack{l_1 \in \{1,\ldots,q+1\} \\ m_1 \in \{1,2,\ldots,T\}}} \{p(k(m_1)_{\ell_1})\}^2 + \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} \{p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2})\}^2 \Bigg]. \tag{28}$$

Finally, we work on bounding $\sum_{i=1}^{n} \Delta_i^3$. Again, by Taylor Expansion, we have

$$\begin{aligned}
\sum_{i=1}^{n} \Delta_i^3 = \sum_{i=1}^{n} \Bigg\{ \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \alpha(k_1,\ldots,k_T) \Bigg[ \sum_{\substack{\ell_1 \in \{1,\ldots,q+1\} \\ m_1 \in \{1,\ldots,T\}}} p(k(m_1)_{\ell_1}) U_{i,\tilde{\beta}_{k(m_1),\ell_1}}(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n) \\
+ \frac{1}{2} \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2}) U_{i,\tilde{\beta}_{k(m_1),\ell_1},\tilde{\beta}_{k(m_2),\ell_2}}(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n) \Bigg] \Bigg\}^3 \\
= O_p(1) \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \Bigg\{ \sum_{\substack{\ell_1 \in \{1,\ldots,q+1\} \\ m_1 \in \{1,\ldots,T\}}} \left| p(k(m_1)_{\ell_1}) \right|^3 \sum_{i=1}^{n} U_{i,\tilde{\beta}_{k(m_1),\ell_1}}^3(\tilde{\boldsymbol{\beta}}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n) \\
+ \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} \left| p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2}) \right|^3 \sum_{i=1}^{n} U_{i,\tilde{\beta}_{k(m_1),\ell_1},\tilde{\beta}_{k(m_2),\ell_2}}^3(\boldsymbol{\xi}^{\{\boldsymbol{k}_t\}}, \tilde{G}_n) \Bigg\} \\
= O_p(n) \sum_{k_1=1}^{K_0} \cdots \sum_{k_T=1}^{K_0} \Bigg\{ \sum_{\substack{\ell_1 \in \{1,\ldots,q+1\} \\ m_1 \in \{1,\ldots,T\}}} \left| p(k(m_1)_{\ell_1}) \right|^3 + \sum_{\substack{\ell_1,\ell_2 \in \{1,\ldots,q+1\} \\ m_1,m_2 \in \{1,\ldots,T\} \\ n_1+n_2=2}} \left| p(k(m_1)_{\ell_1}^{n_1} k(m_2)_{\ell_2}^{n_2}) \right|^3 \Bigg\} = o_p(n) \|\boldsymbol{R}\|_F^2,
\end{aligned}$$

implying $\sum_{i=1}^{n} \Delta_i^2$ dominates $\sum_{i=1}^{n} \Delta_i^3$. Hence, we can rewrite (25) as

$$\ell_n(G,\boldsymbol{\beta}) - \ell_n(\tilde{G}_n,\boldsymbol{\beta}) \leq \sum_{i=1}^{n} \Delta_i - \left[ \frac{1}{2} \sum_{i=1}^{n} \Delta_i^2 \right] (1 + o_p(1)).$$

For large $n$, combining with (27) and (28), for some constant $C$, we simplify this result as

$$\ell_n(G,\boldsymbol{\beta}) - \ell_n(\tilde{G}_n,\boldsymbol{\beta}) \leq \sum_{i=1}^{n} \Delta_i - \frac{1}{2} \sum_{i=1}^{n} \Delta_i^2 \leq C n^{\frac{3}{4}} \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{I}_k} \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_j \right\|_2^2 \lesssim C n^{\frac{1}{2}} \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{I}_k} \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_j \right\|_2.$$

Then together with (24), it eventually leads to

$$\tilde{\ell}_n(G,\boldsymbol{\beta}) - \tilde{\ell}_n(\tilde{G}_n,\boldsymbol{\beta}) \leq C n^{\frac{1}{2}} \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{I}_k} \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_j \right\|_2 - n\lambda_n \sum_{k=1}^{K_0} \sum_{j,j+1 \in \mathcal{I}_k} \left\| \boldsymbol{\eta}_j \right\|_2.$$

Since $\lambda_n = O(n^{-\frac{1}{4}} \log n) \to \infty$, the right-hand side of above inequality tends to be negative for large $n$. It shows any $G$ with strictly $\hat{K}_n > K_0$ can not be a MPLE. The proof of Theorem 2 is completed. □

**Proof of Theorem 3.** This proof technique is extended from [40] and [35]. The extended ITD algorithm updates $\boldsymbol{\eta}$ through

$$\boldsymbol{\eta}^{(m+1)} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\ \tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m)}).$$

To show the decrease of $G(\boldsymbol{\eta})$ iterative, we notice the relation that $G(\boldsymbol{\eta}) = \rho^{-1}\tilde{Q}(\boldsymbol{\eta};\boldsymbol{\eta})$. Thus,

$$G(\boldsymbol{\eta}^{(m)}) = \rho^{-1}\tilde{Q}(\boldsymbol{\eta}^{(m)};\boldsymbol{\eta}^{(m)}) \geq \rho^{-1}\tilde{Q}(\boldsymbol{\eta}^{(m+1)};\boldsymbol{\eta}^{(m)}) \tag{29}$$
$$= G(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}\rho^{-1}\sum_{j=1}^{K}\|\boldsymbol{\eta}_j^{(m+1)} - \boldsymbol{\eta}_j^{(m)}\|_2^2 - \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m+1)}) - b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m)}) - b'(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m)})\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)})\right]\right\}h_{itk}\right].$$

Using Taylor's expansion, for some $\tilde{\boldsymbol{\beta}}_k$ between $\boldsymbol{\beta}_k^{(m+1)}$ and $\boldsymbol{\beta}_k^{(m)}$, we obtain

$$b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m+1)}) - b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m)}) - b'(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^{(m)})\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)})\right] = \frac{1}{2}b''(\boldsymbol{x}_{it}^T\tilde{\boldsymbol{\beta}}_k)\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)})\right]^2.$$

Together with $h_{itk} \in (0,1)$, $\tau_1$ and the definition of $\tau_2^{(m)}$, the following is obtained:

$$\begin{aligned}
G(\boldsymbol{\eta}^{(m)}) &\geq G(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}\rho^{-1}\sum_{j=1}^{K}\|\boldsymbol{\eta}_j^{(m+1)} - \boldsymbol{\eta}_j^{(m)}\|_2^2 - \frac{\tau_2^{(m)}}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)})\right]^2 \\
&\geq G(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}\sum_{j=1}^{K}(\boldsymbol{\eta}_k^{(m+1)} - \boldsymbol{\eta}_k^{(m)})^T[\rho^{-1}\boldsymbol{I} - (K-j+1)\tau_2^{(m)}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{x}_{it}\boldsymbol{x}_{it}^T](\boldsymbol{\eta}_k^{(m+1)} - \boldsymbol{\eta}_k^{(m)}) \\
&\geq G(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}(\rho^{-1} - K\tau_2^{(m)}\tau_1)\sum_{j=1}^{K}\|\boldsymbol{\eta}_j^{(m+1)} - \boldsymbol{\eta}_j^{(m)}\|_2^2,
\end{aligned}$$

because

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)})\right]^2 = \sum_{j=1}^{K}\sum_{k=j}^{K}\left\{(\boldsymbol{\eta}_k^{(m+1)} - \boldsymbol{\eta}_k^{(m)})^T[\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{x}_{it}\boldsymbol{x}_{it}^T](\boldsymbol{\eta}_k^{(m+1)} - \boldsymbol{\eta}_k^{(m)})\right\}.$$

If the condition $\rho^{-1} \geq K\tau_2^{(m)}\tau_1$ holds in theorem 3, then $G(\boldsymbol{\eta}^{(m)}) \geq G(\boldsymbol{\eta}^{(m+1)})$ is guaranteed.

Furthermore, we let $\tau_2^* = \sup_{m\geq 1}\tau_2^{(m)}$ and we have

$$G(\boldsymbol{\eta}^{(m)}) - G(\boldsymbol{\eta}^{(m+1)}) \geq \frac{1}{2}(\rho^{-1} - K\tau_2^*\tau_1)\sum_{j=1}^{K}\|\boldsymbol{\eta}_j^{(m+1)} - \boldsymbol{\eta}_j^{(m)}\|_2^2.$$

Thus, $G(\boldsymbol{\eta}^{\boldsymbol{m}})$ is bounded and non-decreasing and leads to $\|\boldsymbol{\eta}_j^{(m+1)} - \boldsymbol{\eta}_j^{(m)}\|_2^2 \to 0$, $j \in \{1,\ldots,K\}$. The compact condition of $\left\{\boldsymbol{\eta} : G(\boldsymbol{\eta}) \leq G\left(\boldsymbol{\eta}^{(0)}\right)\right\}$ also indicates the existence of limit point $\boldsymbol{\eta}^*$. Suppose $\{\boldsymbol{\eta}^{m_t}\}$ is a subsequence of $\{\boldsymbol{\eta}^m\}$ such that $\boldsymbol{\eta}^{m_t} \to \boldsymbol{\eta}^*$ as $t \to \infty$. According to formula of ITD updating, we have

$$\boldsymbol{\eta}^{(m_t+1)} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\ \tilde{Q}(\boldsymbol{\xi};\boldsymbol{\eta}^{(m_t)})$$

by the bivariate continuity of $\tilde{Q}(\boldsymbol{\xi};\boldsymbol{\eta})$, taking a limit on both sides. Thus, we obtain

$$\begin{aligned}
\boldsymbol{\eta}^* &= \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\ \tilde{Q}(\boldsymbol{\xi};\boldsymbol{\eta}^*) = \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\ \rho G(\boldsymbol{\xi}) + T(\boldsymbol{\xi};\boldsymbol{\eta}^*), \\
T(\boldsymbol{\xi};\boldsymbol{\eta}^*) &= \frac{1}{2}\rho^{-1}\sum_{j=1}^{K}\|\boldsymbol{\eta}_j - \boldsymbol{\eta}_j^*\|_2^2 - \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k) - b(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^*) - b'(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_k^*)\left[\boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^*)\right]\right\}h_{itk}\right].
\end{aligned}$$

We can observe that $\boldsymbol{\eta}^*$ is the local minimum point of $T(\boldsymbol{\xi};\boldsymbol{\eta}^*)$, i.e., $\left.\frac{\partial T(\boldsymbol{\xi},\boldsymbol{\eta}^*)}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\eta}^*} = \mathbf{0}$. Consequently, $\boldsymbol{\eta}^*$ is also the stationary point of $G(\boldsymbol{\eta})$.

Then, to prove the convergence of $\boldsymbol{\eta}^{(m)}$, by using contradiction, we suppose that $\{\boldsymbol{\eta}^{(m)}\}$ does not converge but has two different limiting points, i.e., $\boldsymbol{\eta}^{*1} \neq \boldsymbol{\eta}^{*2}$. Let $\epsilon = \sum_{j=1}^{K} \|\boldsymbol{\eta}_j^{*1} - \boldsymbol{\eta}_j^{*2}\|_2^2$. The discrepancy between successive $\boldsymbol{\eta}^{(m+1)}$ and $\boldsymbol{\eta}^{(m)}$ tends to 0 as $m \to 0$ shows an infinite number of $\boldsymbol{\eta}^{(m)}$ between them. Thus, we can pick up $\{\boldsymbol{\eta}^{(m)} : \sum_{j=1}^{K} \|\boldsymbol{\eta}_j^{(m)} - \boldsymbol{\eta}_j^{*1}\|_2^2 > \epsilon/3$ and $\sum_{j=1}^{K} \|\boldsymbol{\eta}_j^{(m)} - \boldsymbol{\eta}_j^{*2}\|_2^2 > \epsilon/3\}$, which contains an infinite number of elements. Equipped with compact assumption, this set must contain another limiting point, say $\boldsymbol{\eta}^{*3}$. We can continue the construction procedure to produce infinite limiting points and form the contradiction toward the assumption. Thus, we have proven the sequence $\{\boldsymbol{\eta}^{(m)}\}$, which further converges to a stationary point of $G(\boldsymbol{\eta})$. □

**Appendix C: Derivation of ECM formula**

Under the EM algorithm terminology, the underlying state trajectories $\boldsymbol{S} = \{(S_{i1}, \ldots, S_{iT})\}, i \in \{1, \ldots, n\}$ are regarded as latent variables. The ECM algorithm aims to maximize the conditional expectation of the complete-data log-likelihood function, $\ell_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X})$, which is given as follows:

$$\begin{aligned}\ell_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}) &= \ln \mathcal{L}(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}) = \ln f(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Psi}) + \ln f(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\Psi}) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it}|S_{it}, \boldsymbol{x}_{it}, \boldsymbol{\beta}) + \sum_{i=1}^{n} \log \pi_{S_{i1}} + \sum_{i=1}^{n} \sum_{t=2}^{T} \ln P_{S_{i,t-1}, S_{it}}.\end{aligned}$$

Then, the penalized complete-data log-likelihood $\tilde{l}_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X})$ can be written as

$$\tilde{\ell}_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}) = \ell_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}) + C_K \sum_{k=1}^{K} \log \pi_k - n \sum_{k=2}^{K} p_{\lambda_n}(\|\boldsymbol{\eta}_k\|_2). \tag{30}$$

In the E-step, we need to construct the Q-surrogate function $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$ by taking the expectation of the penalized complete-data log-likelihood (30) with respect to $\boldsymbol{S}$ conditional on $\boldsymbol{Y}$ and the $p$th iteration of $\boldsymbol{\Psi}^{(p)}$ as follows:

$$Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)}) = E\left(\tilde{\ell}_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}) \,|\boldsymbol{Y}, \boldsymbol{\Psi}^{(p)}\right) = \sum_{\boldsymbol{S}} \tilde{\ell}_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X})\, f(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}),$$

where $\sum_{\boldsymbol{S}}$ represents the summation of all possible state trajectories for each subject, and

$$f(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}) \propto f(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}) f(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\Psi}^{(p)}).$$

We let $h^{(p)}(\boldsymbol{S}) = f(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}) f(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\Psi}^{(p)})$ and rewrite $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$ in (6) into

$$Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)}) \propto \sum_{\boldsymbol{S}} \tilde{\ell}_n^c(\boldsymbol{\Psi}; \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X})\, h^{(p)}(\boldsymbol{S}) = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\begin{aligned}\mathrm{I} &= \sum_{\boldsymbol{S}} \left( \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(y_{it}|S_{it}, \boldsymbol{x}_{it}, \boldsymbol{\Psi}) \right) h^{(p)}(\boldsymbol{S}) - n \sum_{k=2}^{K} p_{\lambda_n}(\|\boldsymbol{\eta}_k\|_2), \\ \mathrm{II} &= \sum_{\boldsymbol{S}} \left( \sum_{i=1}^{n} \log \pi_{S_{i1}} \right) h^{(p)}(\boldsymbol{S}) + C_K \sum_{k=1}^{K} \log \pi_k, \\ \mathrm{III} &= \sum_{\boldsymbol{S}} \left( \sum_{i=1}^{n} \sum_{t=2}^{T} \log P_{S_{i,t-1}, S_{it}} \right) h^{(p)}(\boldsymbol{S}).\end{aligned} \tag{31}$$

The conditional maximization (CM) step maximizes $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$ with respect to $\boldsymbol{\Psi}$ to obtain

$$\boldsymbol{\Psi}^{(p+1)} = \underset{\boldsymbol{\Psi}}{\operatorname{argmax}}\ Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)}). \tag{32}$$

In the CM step, we split $\boldsymbol{\Psi}$ into two sets of parameters $\boldsymbol{\Psi}_1 = \{\pi_1, \ldots, \pi_K; P_{11}, P_{12}, \ldots, P_{KK}\}$ and $\boldsymbol{\Psi}_2 = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$. Then, maximizing $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$ over $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ sequentially leads to the conditional updating steps as follows:

$$\begin{aligned}
&\text{CM-step 1: } \boldsymbol{\Psi}_1^{(p+1)} = \underset{\boldsymbol{\Psi}_1}{\operatorname{argmax}}\ Q(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2^{(p)}|\boldsymbol{\Psi}^{(p)}),\\
&\text{CM-step 2: } \boldsymbol{\Psi}_2^{(p+1)} = \underset{\boldsymbol{\Psi}_2}{\operatorname{argmax}}\ Q(\boldsymbol{\Psi}_1^{(p+1)}, \boldsymbol{\Psi}_2|\boldsymbol{\Psi}^{(p)}).
\end{aligned}$$

Along with the forward–backward technique in HMM, we denote

$$\begin{aligned}
a_{ik}^{(p)}(t) &= f\left(y_{i1}, \ldots, y_{it}, S_{it} = k|\boldsymbol{X}, \boldsymbol{\Psi}^{(p)}\right),\\
b_{ik}^{(p)}(t) &= f\left(y_{i,t+1}, \ldots, y_{iT}|S_{it} = k, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}\right)
\end{aligned} \tag{33}$$

as forward and backward probabilities, respectively, and they can be calculated in a recursive manner through the Baum–Welch algorithm [4]. By combining (33) with the Markov conditional independence property, we can simplify $h^{(p)}(\boldsymbol{S})$ to a constant into

$$\begin{aligned}
h^{(p)}(S_{it} = k) &= \frac{a_{ik}^{(p)}(t) b_{ik}^{(p)}(t)}{\sum_{h=1}^{K} a_{ih}^{(p)}(T)},\\
h^{(p)}(S_{i,t-1} = r, S_{it} = s) &= \frac{a_{ir}^{(p)}(t-1) P_{rs} f(y_{it}|S_{it} = s, \boldsymbol{X}, \boldsymbol{\Psi}^{(p)}) b_{is}^{(p)}(t)}{\sum_{k=1}^{K} a_{ik}^{(p)}(T)}.
\end{aligned}$$

Employing these formulas, we can derive the closed form of $\boldsymbol{\Psi}_1^{(p+1)}$ in CM step 1 by examining expressions (31) and (32) directly with some basic algebraic manipulations. Specifically, the two CM steps can be rewritten as following:

$$\pi_k^{(p+1)} = \frac{\sum_{i=1}^{n} h^{(p)}(S_{i1} = k) + C_k}{n + KC_k}, \quad P_{rs}^{(p+1)} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T} h^{(p)}(S_{i,t-1} = r, S_{it} = s)}{\sum_{i=1}^{n}\sum_{t=2}^{T} h^{(p)}(S_{i,t-1} = r)}. \tag{34}$$

$$\boldsymbol{\beta}^{(p+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{k=1}^{K} \left[\sum_{i=1}^{n}\sum_{t=1}^{T} \ln f(y_{it}|S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}) h^{(p+1)}(S_{it} = k)\right] - n\sum_{k=2}^{K} p_{\lambda_n}(\|\boldsymbol{\eta}_k\|_2), \tag{35}$$

where $k, r, s \in \{1, \ldots, K\}$ and $\boldsymbol{\eta}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}$.

**Appendix D: Definition of Multivariate Thresholding Operator**

Here, we provide the specific formulation for multivariate thresholding operator $\vec{\boldsymbol{S}}$ as following:

$$\vec{\boldsymbol{S}}(z; \kappa, a, \lambda_n) = z^{\circ} S(\|z\|_2; \kappa, a, \lambda_n) \text{ with } z^{\circ} = \begin{cases} \frac{z}{\|z\|_2}, & \text{if } z \neq \mathbf{0},\\ \mathbf{0}, & \text{if } z = \mathbf{0}, \end{cases}$$

where

- when $0 < \kappa \leq a - 1$,

$$S(\|z\|_2; \kappa, a, \lambda_n) = \begin{cases} (\|z\|_2 - \kappa\lambda_n)_+ \cdot \operatorname{sgn}(\|z\|_2), & \text{when } \|z\|_2 < (\kappa+1)\lambda,\\ \dfrac{(a-1)\|z\|_2 - \kappa a\lambda \operatorname{sgn}(\|z\|_2)}{a - \kappa - 1}, & \text{when } (\kappa+1)\lambda \leq \|z\|_2 < a\lambda,\\ \|z\|_2, & \text{when } \|z\|_2 \geq a\lambda. \end{cases}$$

- when $a - 1 < \kappa \leq a$,

$$S(\|z\|_2; \kappa, a, \lambda_n) = \begin{cases} (\|z\|_2 - \kappa\lambda_n)_+ \cdot \operatorname{sgn}(\|z\|_2), & \text{when } \|z\|_2 < a\lambda_n\\ \|z\|_2, & \text{when } \|z\|_2 \geq a\lambda_n \end{cases}$$

- when $a < \kappa$,

$$S(\|z\|_2; \kappa, a, \lambda_n) = \|z\|_2 I(\|z\|_2 \geq a\lambda_n).$$

## Appendix E: Online Supplementary Material

The online supplementary materials contains pseudo-code table bluein for ECM-ITD algorithm, settings and additional results in Simulations 1 and 2, and additional results in the real data analysis.

## References


[1] A. Agresti, Categorical Data Analysis, volume 482, John Wiley & Sons, New York, 2003.
[2] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.
[3] R. M. Altman, Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting, Journal of the American Statistical Association 102 (2007) 201–210.
[4] L. E. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains, The Annals of Mathematical Statistics 41 (1970) 164–171.
[5] M. J. Beal, Variational algorithms for approximate Bayesian inference, Ph.D. thesis, UCL (University College London), 2003.
[6] P. J. Bickel, Y. Ritov, T. Rydén, Asymptotic normality of the maximum-likelihood estimator for general hidden markov models, The Annals of Statistics 26 (1998) 1614–1635.
[7] J. Chen, Optimal rate of convergence for finite mixture models, The Annals of Statistics (1995) 221–233.
[8] J. Chen, A. Khalili, Order selection in finite mixture models with a nonsmooth penalty, Journal of the American Statistical Association 104 (2009) 187–196.
[9] D. Dacunha-Castelle, E. Gassiat, Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes, The Annals of Statistics 27 (1999) 1178–1209.
[10] B. C. Dickerson, D. Wolk, Biomarker-based prediction of progression in mci: comparison of ad-signature and hippocampal volume with spinal fluid amyloid-$\beta$ and tau, Frontiers in Aging Neuroscience 5 (2013) 55.
[11] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge university press, Cambridge, England, 1998.
[12] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American statistical Association 96 (2001) 1348–1360.
[13] S. Frühwirth-Schnatter, R. Frühwirth, Auxiliary mixture sampling with applications to logistic models, Computational Statistics & Data Analysis 51 (2007) 3509–3528.
[14] M. Goedert, Neurofibrillary pathology of alzheimer's disease and other tauopathies, Progress in Brain Research 117 (1998) 287–306.
[15] Y. Hung, Y. Wang, V. Zarnitsyna, C. Zhu, C. J. Wu, Hidden markov models with applications in cell adhesion experiments, Journal of the American Statistical Association 108 (2013) 1469–1479.
[16] E. Ip, Q. Zhang, J. Rejeski, T. Harris, S. Kritchevsky, Partially ordered mixed hidden markov model for the disablement process of older adults, Journal of the American Statistical Association 108 (2013) 370–384.
[17] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand, et al., Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers, The Lancet Neurology 12 (2013) 207–216.
[18] F. Jelinek, Statistical Methods for Speech Recognition, MIT press, 1997.
[19] F. Jessen, S. Wolfsgruber, B. Wiese, H. Bickel, E. Mösch, H. Kaduszkiewicz, M. Pentzek, S. G. Riedel-Heller, T. Luck, A. Fuchs, et al., Ad dementia risk in late mci, in early mci, and in subjective memory impairment, Alzheimer's & Dementia 10 (2014) 76–83.
[20] K. Kang, J. Cai, X. Song, H. Zhu, Bayesian hidden markov models for delineating the pathology of alzheimers disease, Statistical Methods in Medical Research 28 (2019) 2112–2124.
[21] K. Kantarci, J. L. Gunter, N. Tosakulwong, S. D. Weigand, M. S. Senjem, R. C. Petersen, P. S. Aisen, W. J. Jagust, M. W. Weiner, C. R. Jack Jr, et al., Focal hemosiderin deposits and $\beta$-amyloid load in the adni cohort, Alzheimer's & Dementia 9 (2013) S116–S123.
[22] B. G. Leroux, M. L. Puterman, Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models, Biometrics (1992) 545–558.
[23] H. Liu, X. Song, Bayesian analysis of hidden markov structural equation models with an unknown number of hidden states, Econometrics and Statistics (2020).
[24] D. A. Llano, G. Laforet, V. Devanarayan, Derivation of a new adas-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to alzheimer disease, Alzheimer Disease & Associated Disorders 25 (2011) 73–84.
[25] I. L. MacDonald, W. Zucchini, Hidden Markov and other models for discrete-valued time series, volume 110, CRC Press, New York, 1997.
[26] R. J. MacKAY, Estimating the order of a hidden markov model, Canadian Journal of Statistics 30 (2002) 573–589.
[27] T. Manole, A. Khalili, Estimating the number of components in finite mixture models via the group-sort-fuse procedure, The Annals of Statistics 49 (2021) 3043–3069.
[28] R. C. Mohs, D. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, et al., Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope., Alzheimer Disease and Associated Disorders (1997).
[29] J. L. Podcasy, C. N. Epperson, Considering sex and gender in alzheimer disease and other dementias, Dialogues in Clinical Neuroscience 18 (2016) 437.
[30] S. L. Risacher, S. Kim, K. Nho, T. Foroud, L. Shen, R. C. Petersen, C. R. Jack Jr, L. A. Beckett, P. S. Aisen, R. A. Koeppe, et al., Apoe effect on alzheimer's disease biomarkers in older adults with significant memory concern, Alzheimer's & Dementia 11 (2015) 1417–1429.
[31] S. L. Risacher, S. Kim, L. Shen, K. Nho, T. Foroud, R. C. Green, R. C. Petersen, C. R. Jack Jr, P. S. Aisen, R. A. Koeppe, et al., The role of apolipoprotein e (apoe) genotype in early mild cognitive impairment (e-mci), Frontiers in Aging Neuroscience 5 (2013) 11.
[32] G. Schwarz, et al., Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461–464.

[33] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, et al., mgene: accurate svm-based gene finding with an application to nematode genomes, Genome Research 19 (2009) 2133–2143.
[34] R. J. Serfling, Approximation theorems of mathematical statistics, volume 162, John Wiley & Sons, New York, 2009.
[35] Y. She, An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors, Computational Statistics & Data Analysis 56 (2012) 2976–2990.
[36] X. Song, Y. Xia, H. Zhu, Hidden markov latent variable models with multivariate longitudinal data, Biometrics 73 (2017) 313–323.
[37] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B 58 (1996) 267–288.
[38] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, Y. Amirat, An unsupervised approach for automatic activity recognition based on hidden markov model regression, IEEE Transactions on Automation Science and Engineering 10 (2013) 829–835.
[39] H. Wang, B. Li, C. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, Journal of the Royal Statistical Society: Series B 71 (2009) 671–683.
[40] C. Xu, J. Chen, A thresholding algorithm for order selection in finite mixture models, Communications in Statistics-Simulation and Computation 44 (2015) 433–453.
[41] D. Yu, L. Deng, Automatic Speech Recognition., Springer, New York, 2016.
[42] C.-H. Zhang, et al., Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics 38 (2010) 894–942.
[43] J. Zhou, X. Song, L. Sun, Continuous time hidden markov model for longitudinal data, Journal of Multivariate Analysis (2020) 104646.
[44] H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429.
[45] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, The Annals of statistics 36 (2008) 1509.
[46] W. Zucchini, I. L. MacDonald, Hidden Markov models for time series: an introduction using R, Chapman and Hall/CRC, New York, 2009.

Author statement:

**Yiqi Lin:** Data curation, Investigation, Methodology, Software, Writing- Original draft preparation, Visualization. **Xinyuan Song:** Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing, Validation