ON MARKOV AND HIDDEN MARKOV MODELS WITH APPLICATIONS TO TRAJECTORIES

by

Jieyu Fan

B.S. in Statistics, Sun Yat-sen University, 2007

M.S. in Mathematical Statistics, Renmin University of China, 2010

Submitted to the Graduate Faculty of

the Department of Statistics, Dietrich Graduate School of Arts and

Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

DIETRICH GRADUATE SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Jieyu Fan

It was defended on

October 6th 2014

and approved by

Satish Iyengar, Department of Statistics, University of Pittsburgh

Boris Birmaher, Department of Psychiatry, University of Pittsburgh

Henry Block, Department of Statistics, University of Pittsburgh

Yu Cheng, Department of Statistics, University of Pittsburgh

Jing Lei, Department of Statistics, Carnegie Mellon University

Dissertation Director: Satish Iyengar, Department of Statistics, University of Pittsburgh

ON MARKOV AND HIDDEN MARKOV MODELS WITH APPLICATIONS TO TRAJECTORIES

Jieyu Fan, PhD

University of Pittsburgh, 2014

Markov and hidden Markov models (HMMs) provide a special angle to characterize trajectories using their state transition patterns. Distinct from Markov models, HMMs assume that an unobserved sequence governs the observed sequence and the Markovian property is imposed on the hidden chain rather than the observed one. In the first part of this dissertation, we develop a model for HMMs with exponential family distribution and extend it to incorporate covariates. We call it HMM-GLM, for which we propose a joint model selection method. The proposed selection criterion is tailored for HMM-GLM aiming at a more accurate approximation of the Kullback-Leibler divergence; we seek improvement of the widely-used Akaike information criterion. The second and the third parts of this dissertation are about clustering trajectories with HMMs and Markov mixture models. The research interests for HMM clustering are to develop a less computationally expensive and more interpretable algorithm for HMM sequence clustering problem, based on the emission and transition features of the chains. We propose an efficient clustering method using Bhattacharyya affinity to measure the pairwise similarity between sequences and apply a spectral clustering algorithm to obtain the cluster assignment. The computational efficiency benefits from the fact that our method avoids iterative computation for the affinity of a pair of sequences. Meanwhile, both simulation and empirical studies show that the proposed algorithm maintains good performance compared to other similar methods. In the third part of the dissertation, we address a study of the course of children and adolescents with bipolar disorder. Measuring and making sense of the fluctuations in different moods over

time is challenging. We use a Markov mixture model with different transition matrices to find homogeneous clusters and capture different longitudinal mood change patterns. We also conduct a simulation study to investigate the performance of the model when there are violations of model assumptions. The results show that this model is fairly robust in the tested situations. We find that the clusters separate out those who tend to stay in a mood state from those who fluctuate between mood states more frequently.

Keywords: hidden Markov model, Markov mixture model, clustering, model selection.

TABLE OF CONTENTS

1.0 INTRODUCTION 1

2.0 JOINT MODEL SELECTION FOR HIDDEN MARKOV MODELS WITH EXPONENTIAL FAMILY OBSERVATIONS 5

2.1 Overview 5

2.2 Model Development 8

2.3 Parameter Estimation 10

2.4 Approximating the Kullback-Leibler Divergence 14

2.5 Simulation 20

2.6 Discussion 24

3.0 CLUSTERING SEQUENCES WITH HIDDEN MARKOV MODELS 25

3.1 Introduction of Hidden Markov Clustering 25

3.2 Method Development 28

3.3 Simulation 30

3.4 Empirical Study: Australian Sign Language 31

3.5 Discussion 32

4.0 CLUSTERING TRAJECTORIES WITH FIRST-ORDER MARKOV MIXTURE MODELS 36

4.1 Introduction of Markov Mixture Models and Bipolar Disorder Study 36

4.2 Model Development 38

4.3 Simulation 40

4.4 Application to Bipolar Youth Study 43

4.5 Discussion 50

BIBLIOGRAPHY 54

LIST OF TABLES

1 Model selection criterion comparison 1 21
2 Model selection criterion comparison 2 22
3 Model selection criterion comparison 3 23
4 Clustering algorithms comparison on Australian sign language data 32
5 Simulation study for Markov mixture model 45
6 Mood rating scale in bipolar youth study 47
7 The estimated transition matrix in bipolar youth study 53

LIST OF FIGURES

1 Australian sign language data sample 1 33
2 Australian sign language data sample 2 34
3 A simulation example 44
4 Example sequences (12-point rating score) 48
5 Example sequences (4-point rating score) 48
6 Weekly mood rating of 412 children and adolescents with bipolar disorder. . . 51
7 Markov mixture model separates three clusters in bipolar youth study 52
8 Splitting the largest cluster in bipolar youth study 52

1.0 INTRODUCTION

Benefiting from modern computational technology, people are able to collect and process more data nowadays. As a result, sequential data is expected to appear more often, which raises the need to develop methodologies to analyze them. The Markov and hidden Markov models are two popular methods for modeling sequential observations. In the exploratory stage, clustering sequences could identify similar sequences and reduce heterogeneity for further study. In particular, clustering with hidden Markov models or Markov models provide a way to group sequences based on the transition patterns.

Markov models are used to model stochastic processes which have the Markovian property that the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state, not on the sequence of events that preceded it. There are two additional properties often assumed for Markov chains: time homogeneity and stationarity. The first says the state transition matrix of the same order does not change over time, while the second says that the distribution of the states becomes stable as time goes by. With these assumptions, a Markov chain is determined by the transition matrix and the initial distribution. We will repeatedly see the benefits of having these properties in our study below.

Different from Markov models, hidden Markov models (HMMs) assume there is an underlying unobserved sequence which governs the observed sequence and that the Markovian property is imposed on the hidden chain instead of the observed one. The most well-known applications of HMMs are in speech recognition. It has since been introduced successfully to other fields, such as bioinformatics. The conditional dependence of the observed sequence and the Markovian properties of the hidden sequence are essential to factor the likelihood function so that it is tractable. In addition, the ability of HMMs to handle both single and

multiple sequences and unequally spaced observations of varying lengths also makes them appealing in many applications. From both practical and theoretical considerations, it is worth more effort and attention to study the model selection problem for HMMs since it is widely used.

Now that we have the models at hand to describe a data sequence, it is of further interest to know whether we can separate out those sequences which evolve in different patterns, which leads to a clustering problem. After clustering, we gain more homogeneous subgroups which we can study further. There is an extensive literature on clustering over the years. Besides traditional clustering methods, such as the K-means algorithm and hierarchical clustering, spectral clustering has recently become more prominent, because it is easy to implement and has shown good performance in practice. However, clustering sequences would be more challenging due to the dependence within the sequential data. One of the challenges lies in finding an appropriate distance measure between sequences. A semi-parametric method we consider in this study is to begin with a sequential model to extract the features of the data and then construct the distance matrix based on the features and apply existing clustering algorithms.

The research aims of this thesis are model selection of HMMs and clustering sequences with Markov models and HMMs. This thesis consists of three parts. In the first part, we consider HMMs with exponential family distribution and extend it to incorporate covariates. Because the way we include the covariates and estimate the coefficients are similar to generalized linear model (GLM), we call the combined model HMM-GLM. We are interested in HMM-GLM and its model selection because even though there are many applications of HMM in various research areas, the literature provides few systematic studies of HMM-GLM and consider the criterion to choose the optimal number of hidden states and variables for it. In this study, we propose a joint model selection method for HMM-GLM. The criterion is considered from a more accurate approximation of the Kullback-Leibler divergence to establish a general criterion like Akaike information criterion (AIC) for HMMs. Our simulation study shows that the proposed criterion is competitive when the number of observations in the sequence is small to medium, which is generally more difficult compared to large sample situations.

The second part of the study is about clustering sequences with HMMs. Sequence clustering is a special kind of clustering problem where the input do not consist of independent observations. The dependent data structure violates model assumptions in many existing clustering methods leading to challenges for researchers. The aim of our study is to develop a less computationally expensive and easy-to-interpret algorithm for the HMM sequence clustering problem, based on the emission and transition features of the chains. Our intuition is that when the HMM sequences can be well represented by their emission probabilities and transition matrices, these two main features can then be used to cluster the observed sequences. We propose an efficient clustering method with HMMs using Bhattacharyya affinity to measure the pairwise similarity between sequences, then apply a spectral clustering algorithm to obtain the cluster assignment. The improvement in efficiency is that we avoid iterative computation for the affinity of a pair of sequences. We show that the iterative computation of the affinity measure can be approximated by a function of the emission distribution and transition matrix. The main progress is made in finding an efficient way to obtain the affinity matrix. Though the methodology should be applicable to general emission distributions, in this study we focus on the exponential family cases since their Bhattacharyya affinity formulas are amenable to analysis because they are available in closed form.

In the third part, we study an alternative clustering method, which is the Markov mixture model (MxM), and apply it to a psychiatric study of the course of children and adolescents with bipolar disorder. Bipolar disorder is characterized by recurrent mood episodes ranging from depression to extreme happiness or irritability. Measuring and making sense of the fluctuations in these moods over time is challenging. To find homogeneous clusters and capture different longitudinal mood change patterns we use a Markov mixture model with different transition matrices. We estimate the parameters of this model using EM algorithm. Further, we conduct a simulation study to investigate the performance of the model when there are violations of model assumptions. The result shows that this model is fairly robust even when certain model assumptions fail. In the application, based on clinical considerations we focus on four mood states: well (formally known as euthymia), mania/hypomania, depression and mixed (a combination of symptoms of mania/hypomania and depression).

Specifically, we are interested in the frequency and patterns of changes among them. We find that the clusters separate out those who tend to stay in a mood state from those who fluctuate between mood states more frequently. In fact, both clustering methods (HMMs and MxMs) have the Markovian assumption, one on the hidden chains, the other on the observed chains. Compared to traditional sequence clustering methods, the Markov models provide a novel angle to characterize the sequences with transition patterns. The reason we adopt MxMs in the bipolar study but not HMMs is data driven. When the number of observed states is small and their interpretation is clear, adding a layer of additional hidden states is not compelling.

2.0 JOINT MODEL SELECTION FOR HIDDEN MARKOV MODELS WITH EXPONENTIAL FAMILY OBSERVATIONS

2.1 OVERVIEW

Hidden Markov models (HMMs) are used to model dynamic systems in which the observed sequences are governed by the underlying hidden Markov chains. The basic estimation procedure for HMMs was established in the 1960s and 1970s. It consists of three algorithms: the Baum-Welch, forward-backward, and Viterbi algorithms. The most well-known and successful application of HMMs is in speech recognition. Rabiner gave a comprehensive tutorial on HMM in speech recognition (Rabiner, 1989). Since then, HMMs have gradually appeared more in other fields, such as bioinformatics (Soding, 2004; Krogh et al. 2001), neuroscience (Camproux et al. 1996), and finance (Mamon and Elliott, 2010). Meanwhile, researchers also studied the statistical properties (Bickel et al., 1996, 1998) and developed more efficient algorithms for HMMs (Gales et al. 1992, Bilmes et al. 1998).

HMMs have a neat and intuitive model structure. The conditional dependence of the observed sequence and the Markovian properties of the hidden sequence are essential to factor the likelihood function. In addition, HMMs have the flexibility to model both single and multiple sequences and handle unequally spaced observations of varying lengths.

Another advantage of HMMs is that they can incorporate time-varying explanatory variables. In general, there are two ways to impose the covariate effects in HMMs: one way allows the covariates affect the emission probability, the other way assume the covariates affect the transition matrices. For instance, in a faecal coliform counts study, Turner et al. (1998) developed a model to superimpose the two-way design to HMMs in a generalized linear model (GLM) framework in which the hidden state affected the intercept of the log-link

function. In another longitudinal health status study, Scott et al. (2005) applied HMMs to learn about the health status switching patterns over time and its association with treatments. They developed an inhomogeneous HMM by introducing the Dirichlet distribution with parameters embedded in a Bayesian hierarchical model, in which different transition probabilities may apply for each observation period. The model for the emission was the multivariate t distribution to handle the heavy tail. Besides fixed effects, random effects can also be taken into account. Altman (2007) proposed a Mixed HMM (MHMM) to incorporate covariates and random effects into HMM with an exponential family distribution for the emission distribution. The random effects allowed for long-term dependence within each sequence. The main difference between MHMM and HMMs is that MHMM assumes the observations in a sequence are no longer independent given only the hidden state but not the random effect.

The most frequently used emission probability in the applications of HMMs with explanatory variables are Gaussian and Poisson distributions. However, to the best of our knowledge, no general form of parameter estimation procedure has been given for exponential family. Thus, before looking into the model selection problem we first provide a systematic model development for HMMs with exponential family distribution as the emission distribution. We adapt the estimation procedures in GLM to the HMM framework. When it comes to the model selection for HMMs, we consider both choosing the optimal number of predictors and the optimal number of hidden states. The most straightforward method to compare model performance is to compare the accuracy. For example, in speech recognition the main interest is the estimation of the hidden states, which is what words are pronounced. In this case, we can assess the model performance by the empirical success rate. The limitation of this way is that it does not provide a sufficient insight of the potential problems in the model.

Apart from accuracy, there are three types of method which could be used to evaluate HMM model performance. The first type is to use model fitness criterions, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). This is the most popular approach since it is easy to obtain. Aguirre-Hernandez and Farewell (2002) proposed a Pearson-type goodness-of-fit test for Markov regression models. In their models, the linear

combination of covariates affects only the logarithm of the transition rates, but not the emission probability. In addition, in their study the states are ordinal and the state transitions are limited to adjacent states. Titman and Shaples (2008) generalized it to HMMs with an absorbing state. However, as the number of explanatory variables increases, especially for continuous variables, how to group the levels for each variable becomes more challenging and could be misleading. Visser et al. (2002) discussed both model selection and goodness-of-fit test for HMMs. They pointed out that when comparing HMMs with different numbers of hidden states, a model with fewer states need not be nested in the model with more states. Hence the likelihood ratio test is not suitable to compare the two for this purpose. In comparison, AIC and BIC are easy to obtain and do not have this limitation. As for the model assessment, they introduced a Pearson χ^2 as the prediction error measure. Smith et al. (2006) proposed an information criterion based on the Kullback-Leibler divergence and showed that the new criterion had nice asymptotic properties. Their criterion consists of two parts: one is a log-likelihood measure of the lack-of-fit, and the other is a penalty for the redundant states and variables. This criterion attempts to balance fitness and parsimony. The limitation is that it was deduced for Gaussian emission.

The second type is to use visualization tools, such as residual plots and QQ plots. Altman (2004) provided a visualization method, namely, plots of the estimated distribution against the empirical distribution to check the lack of fit of HMMs with large samples. The idea is based on the fact that as the sequence becomes longer, the empirical distribution converges to the true distribution under certain conditions. However, this conclusion is based on a strong assumption that the observed sequence is strictly stationary, which rarely holds in practice. Zucchini and Macdonald (2009) summarized several existing methods to evaluate HMM fitting. In summary, they suggested the use of AIC and BIC to decide the number of hidden states and then to use pseudo-residual plots as supplements to assess model adequacy.

The third type is to integrate the model selection into the algorithm, treating the number of hidden states as a unknown parameter to estimate with other model parameters. Johansson et al. (2007) developed a Bayesian model selection method for choosing the best number of hidden states for discrete HMMs. They approximated the posterior probability of the hidden sequence with an estimated transition matrix and compared the different model

hypotheses. Recently, Scott et al. (2012) adopted Chib's method (1995, 2001) and BIC to choose the number of hidden states. The former one estimates the posterior distribution of the hidden sequence given the observed sequence from the MCMC output, which was notoriously difficult to calculate. Zhu et al. (2012) proposed a simultaneous model selection procedure for choosing the optimal number of covariates and hidden states in continuous HMMs using the variational Bayesian (VB) algorithm. These integrative algorithm methods usually consider model selection for HMMs from the Bayesian perspective. However, they impose more complicated model structures on the data. As a result, more assumptions are needed and the algorithms become more computationally intensive.

In this study, we aim to develop a joint model selection criterion to simultaneously indicate the optimal number of hidden states and the optimal number of variables for HMMs. We consider to include explanatory variables in the emission probability. Our method is designed for HMMs with exponential family distribution, which is applicable for many applications. Our work is based on the HMM-GLM setting and we develop the model selection index by approximating the Kullback-Leibler divergence. The simulation study shows that the proposed criterion works better than AIC and BIC for small to medium sample sizes, which are generally more challenging.

The organization of this chapter is as follows: In section 2.2, we specify the model setting. In section 2.3, we provide the parameter estimation procedure, especially for how to obtain estimations for coefficients of the explanatory variables. In section 2.4, we show the development of the proposed model selection criterion based on the given model settings. In section 2.5, we present the simulation study. We conclude with a discussion in section 2.6.

2.2 MODEL DEVELOPMENT

Suppose there is an unobserved first-order Markov chain $\{X_t\}_{t=1}^T$ defined on a finite state space ($\leq K$ states) and a corresponding sequence of scalar observations $\{Y_t\}_{t=1}^T$. Assume that the hidden Markov chain is homogeneous and stationary; thus the hidden Markov chain is determined by the initial probability $\pi = (\pi_1, \ldots, \pi_K)$ with constraint $\sum_{k=1}^K \pi_k = 1$ ($\pi_k \geq 0$)

and $K \times K$ transition matrix $A = \{a_{ij}\}$. (The stationarity assumption is not necessary; an ergodic Markov chain is enough to proceed the estimation procedures in HMMs.)

Here we define

$$\pi_k = P(X_1 = k) \tag{2.1}$$

$$a_{ij} = P(X_{t+1} = j | X_t = i) \tag{2.2}$$

$$P(\mathbf{X}; \pi, A) = \pi_{x_1} \prod_{t=1}^{T-1} a_{x_t, x_{t+1}} \tag{2.3}$$

where $X = (X_1, \ldots, X_T)$.

We consider an exponential family for the conditional distribution of the observations (also called emission distribution) in this study. Let $f(y_t|x_t)$ denote the conditional distribution of $Y|X$ at time t, and ψ represent a collection of parameters in the emission distribution. The density function of exponential family distribution is given below, following McCulloch and Searle (2001).

$$f(y_t|x_t; \psi) = \exp\left[\frac{(\gamma_{x_t} y_t - b(\gamma_{x_t}))}{\tau^2} + c(y_t, \tau^2)\right] \tag{2.4}$$

To include time-varying explanatory variables Z ($Z_t \in \mathcal{R}^D$, and $z_t^{'}$ is row t of matrix Z) in the HMM setting, the generalized linear model (GLM) setting can be adopted with a link function $g(\mu_t) = z_t^{'}\beta_{x_t}$. A special kind of link function is called canonical link function as $g(\mu_t) = \gamma_{x_t} = z_t^{'}\beta_{x_t}$. Here, $\mu_t = E(Y_t|X_t)$ and β_{x_t} is a constant coefficient vector given a hidden state. This model setting allows the marginal effects of the explanatory variables to change with the hidden state.

Let $S_t = (S_{t,1}, \ldots, S_{t,K})$ be a K dimensional vector with $S_{t,k} = 1$ when $X_t = k$ and 0 otherwise, and let B be a $D \times K$ coefficient matrix. We can then re-write the canonical link function in matrix form

$$g(\mu_t) = \gamma_{x_t} = z_t^{'} B S_t \tag{2.5}$$

with all entries in B constant. We will also consider non-canonical link functions in case the canonical ones cannot adequately fit the data.

2.3 PARAMETER ESTIMATION

The estimation procedure for the above model can be implemented by combining the estimation techniques of GLM and HMM (HMM-GLM hereafter). The framework of HMMs parameter estimation remains the same as the classical work summarized by Rabiner (1989):
(i) OBSERVED SEQUENCE: How do we estimate the probability of the observed the sequence $\{Y_t\}_{t=1}^T$ given the parameter $\theta = (\pi, A, \psi) = (\pi, A, B, \tau^2)$?
(ii) HIDDEN STATES: Given the parameters, how can we find the conditional distribution of $(X_t|Y_1, \ldots, Y_t)$ (filtering); $(X_t|Y_1, \ldots, Y_T)$ for $T > t$ (smoothing); $(X_t|Y_1, \ldots, Y_s)$ for $s < t$ (prediction)?
(iii) PARAMETER ESTIMATION: How to estimate the parameters θ?

The GLM is relevant in the third part. Later we will see that (i) is related to the likelihood function and construction of the Kullback-Leibler divergence, while (ii) serves the main purpose of HMM, which is to extract information about the hidden sequence from the observed data. These two questions are addressed by the forward-backward and the Viterbi algorithms, which are well-established. The standard approach for question (iii) was proposed by Baum and Welch in 1960s, which is essentially an EM algorithm treating the hidden sequence as missing. Bickel, et al. (1998) showed that the maximum likelihood estimate (MLE) for HMM has an asymptotic normal distribution provided the following conditions hold: the Markov chain is stationary; the expectation of the second derivatives of the conditional distribution (exponential family) density with respect to the parameters exists. Also, with an appropriate starting point, the numerical estimates given by EM tends to the MLE as the number of iterations of EM tends to infinity. We provide more details in answering the last question since we are interested in the effects of the covariates under different hidden states.

The dummy variable S_t of the hidden state X_t is useful here. The complete log-likelihood

for the E-step is

$$\begin{aligned}
\log L(\theta; y, x) &= \sum_{t=1}^{T} \log f(y_t|x_t) + \sum_{t=1}^{T-1} \log(a_{x_t, x_{t+1}}) + \log(\pi_{x_1}) \\
&= \sum_{t=1}^{T} \sum_{k=1}^{K} s_{t,k} \log f(y_t|s_{t,k}) + \sum_{t=1}^{T-1} \sum_{k=1}^{K} \sum_{j=1}^{K} s_{t,k} s_{t+1,j} \log(a_{kj}) + \sum_{k=1}^{K} s_{1,k} \pi_k \\
&= \sum_{t=1}^{T} S_t' \log(f_t) + \sum_{t=1}^{T-1} S_t' [\log(A)] S_{t+1} + S_1' \pi, \qquad (2.6)
\end{aligned}$$

where $f_t = [f(y_t|s_{t,1}), \ldots, f(y_t|s_{t,K})]$. The conditional expectation with respect to the hidden state is

$$\begin{aligned}
Q(\theta, \theta^{old}) &= E_{\theta^{old}}\{\log L(\theta; y_1, \ldots, y_T, x_1, \ldots, x_T)|Y\} \\
&= \sum_{t=1}^{T} \sum_{k=1}^{K} E_{\theta^{old}}(S_{t,k}|Y) \log f(y_t|S_{t,k}; B, \tau^2) \\
&+ \sum_{t=1}^{T-1} \sum_{k=1}^{K} \sum_{j=1}^{K} E_{\theta^{old}}(S_{t,k} S_{t+1,j}|Y) \log(a_{kj}; A) + \sum_{k=1}^{K} E_{\theta^{old}}(S_{1,k}|Y) \pi_k.
\end{aligned} \qquad (2.7)$$

Since $S_{t,k}$ takes on value $\{0, 1\}$, $E(S_{t,k}|Y)$ equals to $P(S_{t,k} = 1|Y)$ which is the smoothing problem in (ii). $E(S_{t,k}, S_{t+1,j}|Y)$ is similar. Denote

$$\zeta_{t,kj}^{old} = E(S_{t,k}, S_{t+1,j}|Y; \theta^{old}) \quad \text{and} \quad \xi_{t,k}^{old} = E(S_{t,k}|Y; \theta^{old}) = \sum_{j=1}^{K} \zeta_{kj}^{old}. \qquad (2.8)$$

We proceed with the estimation assuming that we know $\zeta_{t,kj}$ and $\xi_{t,k}$, which are addressed by the forward-backward algorithm in question (ii). Notice that maximizing $Q(\cdot)$ with respect to (B, τ^2) does not involve the updated (A, π) but only the old (A, π), which is convenient. That is to say no matter what emission distribution is assumed (not limited to the exponential family), it will not affect the algorithms for the estimation of the transition matrices and initial distribution of the hidden Markov chain. Thus the classic forward-backward algorithm and Viterbi algorithm for HMMs are directly applicable.

In the M-step, we need to choose θ to maximize $Q(\theta^{old}, \theta)$. We rewrite the Q-step log-likelihood function with $(\zeta_{kj}^{old}, \xi_k^{old})$

$$\begin{aligned} Q(\theta, \theta^{old}) &= \sum_{t=1}^{T}\sum_{k=1}^{K} \xi_{t,k}^{old}\left[\frac{(z_t'\beta_k y_t - b(z_t'\beta_k))}{\tau^2} + c(y_t, \tau^2)\right] \\ &\quad + \sum_{t=1}^{T-1}\sum_{k=1}^{K}\sum_{j=1}^{K} \zeta_{t,kj}^{old} \log(a_{kj}; A) + \sum_{k=1}^{K} \xi_{t,k}^{old}\pi_k. \end{aligned} \tag{2.9}$$

Here we only show how to estimate B. Note that

$$\max_B Q(\theta, \theta^{old}) = \max_B \sum_{t=1}^{T}\sum_{k=1}^{K} \xi_{t,k}^{old} \frac{(z_t'BS_t y_t - b(z_t'BS_t))}{\tau^2} \tag{2.10}$$

is very similar to GLM except for the "weighting" parameter $\xi_{t,k}$. Hence the procedure for parameter estimation of the GLM can be used. By the chain rule and the relation $\frac{\partial b(\gamma_{t,k})}{\partial \gamma_{t,k}} = \mu_{t,k}$, we get the likelihood equations as

$$Z'W_k^{old}Y = Z'W_k^{old}\mu_k \tag{2.11}$$

where $\mu_{t,k} = g^{-1}(z_t'\beta_k)$, $\mu_k = (\mu_{1,k}, \ldots, \mu_{T,k})$, $Z = (Z_1, \ldots, Z_T)'$ and $W_k = diag(\xi_{1,k}^{old}, \ldots, \xi_{T,k}^{old})$. Here Z is a $T \times D$ matrix. Furthermore, if stationarity holds, the equation reduces to $Z'Y = Z'\mu_k$, because $ES_{t,k} = ES_{t',k}$ for any $t, t' > 0$. Iterative procedures like Newton-Raphson or Fisher's Scoring method can be applied directly for estimating β.

The details are similar for non-canonical links (McCulloch, et al., 2001). The likelihood equations are

$$Z'W_kQ_kG_kY = Z'W_kQ_kG_k\mu_k \tag{2.12}$$

where $Q_k = diag(q_{1,k}, \ldots, q_{T,k})$, $q_{t,k} = [v(\mu_{t,k})g_\mu^2(\mu_{t,k})]^{-1}$, $v(\mu_{t,k}) = Var(Y_t)/\tau^2$ is the variance function, $G_k = diag(g_\mu(\mu_{1,k}), \ldots, g_\mu(\mu_{T,k}))$ and $g_\mu = \partial g/\partial\mu$. The updating formula using Fisher scoring is

$$\beta_k^{(l+1)} = \beta_k^{(l)} + (Z'W_k^{(l)}Q_k^{(l)}Z)^{-1}Z'W_k^{(l)}Q_k^{(l)}G_k^{(l)}(Y - \mu_k^{(l)}). \tag{2.13}$$

where W_k, Q_k, G_k and μ_k are evaluated at $\beta_k^{(l)}$.

The nuisance parameter τ^2 is estimated via a moment estimator based on its relation to the variance of Y (McCullagh and Nelder, p. 328),

$$\hat{\tau}^2_{GLM} = \frac{1}{N-D}\sum_{n=1}^{N}\frac{(y_n-\hat{\mu}_n)^2}{v(\hat{\mu}_n)},$$

which can be modified thus to our case:

$$\hat{\tau}^2 = \frac{1}{T-D}\sum_{i=1}^{K}\sum_{t=1}^{T}\frac{(y_t-\hat{\mu}_{t,k})^2}{v(\hat{\mu}_{t,k})}\hat{\xi}_{t,k}. \tag{2.14}$$

The reason we adopt the estimator from the independent case to our model GLM-HMM is the conditional independence assumption of HMMs, which says given X_t, Y_t is independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_T$.

So far, we have discussed the parameter estimation for the emission probability $f_{y|x}$. Estimation of A and π is similar except for adding Lagrange multipliers for the constraint $\sum_{j=1}^{K} a_{ij} = 1$, $\sum_{k=1}^{K}\pi_k = 1$ and $0 \le a_{ij}, \pi_k \le 1$ in likelihood function. In particular, analogous calculations give

$$a^{new}_{k,j} = \frac{\sum_{t=1}^{T-1}\zeta^{old}_{tkj}}{\sum_{t=1}^{T-1}\xi^{old}_{tk}}. \tag{2.15}$$

If we assume a stationary Markov chain, the initial distribution π is determined by the transition matrix and can be solved as the left eigenvector of A.

Finally, we summarize how to obtain (ζ_{tkj}, ξ_{tk}) by the forward-backward algorithm. Define a forward variable $\alpha_{t,k}$ as

$$\alpha_{t,k} = P(Y_1, \ldots, Y_t, X_t = k) = \left[\sum_{i=1}^{K}\alpha_{t-1,i}a_{ik}\right] f(Y_t|X_t = k), \tag{2.16}$$

and backward variable $b_{t,k}$ as

$$b_{t,k} = P(Y_{t+1}, \ldots, Y_T|X_t = k) = \sum_{i=1}^{K} a_{ki} f(Y_{t+1}|X_{t+1} = i) b_{t+1,i}, \tag{2.17}$$

so that

$$\zeta^{(l)}_{tkj} = \frac{\alpha^{(l)}_{tk}a^{(l)}_{kj}f(y_{t+1}|x_{t+1}=k)b^{(l)}_{t+1,j}}{\sum_{k=1}^{K}\sum_{j=1}^{K}\alpha^{(l)}_{tk}a^{(l)}_{kj}f(y_{t+1}|x_{t+1}=k)b^{(l)}_{t+1,j}} \tag{2.18}$$

$$\xi^{(l)}_{tk} = \sum_{j=1}^{K}\zeta^{(l)}_{tkj} \tag{2.19}$$

for the lth iteration. For more details, see Rabiner et al. (1989).

2.4 APPROXIMATING THE KULLBACK-LEIBLER DIVERGENCE

Suppose that the true underlying model is L_0 and L_1 is the proposed HMM. Followed what Hurvich and Tsai (1989) did in their paper, we assume that the true model is an HMM. The Kullback-Leibler divergence (D_{KL}) measures the information gain (loss) between L_0 and L_1 by taking the expectation of the log-likelihood difference. AIC is also an approximation of the Kullback-Leibler divergence.

The Kullback-Leibler divergence of L_0 and L_1 is defined as:

$$D_{KL}(L_0, L_1) = E_{L_0}\left[\log \frac{L_0(Y)}{L_1(Y)}\right] \tag{2.20}$$

It is nonnegative but need not be symmetric.

The Kullback-Leibler divergence is often used for model comparison. Unlike log-likelihood ratio test, this model comparison does not assume that one model nested in another. And for model comparison, only the log-likelihood of the proposed model matters, since the log-likelihood of the true model will not change in D_{KL}. Thus, let us define the core part in D_{KL} for a candidate model L_1 as

$$D_c(L_1) = -E_Y \log\left[L_1(Y|\hat{\theta}_1)\right].$$

where θ_1 is a set of parameters in model L_1.

Consider the case where the emission distributions are from exponential family. In the approximation, we distinguish training data Y and evaluation data Y^* to avoid underestimation of error because of using the same data. The parameter estimators are functions of the training data Y. We approximate the expectation of D_c by first taking expectation with respect to Y^*, then to Y. Let L_1 be the proposed HMM, then

$$\begin{aligned} D_c(L_1(Y)) &= -E_Y E_{Y^*} \log\left[L_1(\hat{\theta}(Y); Y^*)\right] \\ &= -E_Y E_{Y^*,S^*}[\log f_{Y|S}(Y^*|S^*; \hat{\psi}(Y)) \\ &\quad + \log f_S(S^*; \hat{A}(Y)) - \log f_{S|Y}(S^*|Y^*; \hat{\theta}(Y))] \end{aligned} \tag{2.21}$$

We focus on the approximation of the first term $f_{Y|S}$, since it directly involves the covariates. Let $\hat{\beta}$ be the MLE of β. Based on the fact that under mild conditions $\hat{\beta}$ converges to β as the sample size increases, we approximate $b(\hat{\beta}_k)$ on β_k with Taylor expansion.

$$b(Z_t'\hat{\beta}_k) \approx b(Z_t'\beta_k) + Z_t'\mu(Z_t'\beta_k)(\hat{\beta}_k - \beta_k) + 1/2(\hat{\beta}_k - \beta_k)'H_t(\beta_k)(\hat{\beta}_k - \beta_k) \tag{2.22}$$

where $H_t(\beta_k) = Z_t v(Z_t'\beta_k)Z_t'$, $\partial b(\gamma)/\partial\gamma = \mu(\gamma)$, $\partial^2 b(\gamma)/\partial\gamma^2 = v(\gamma)$. So

$$\begin{aligned} & Z_t'\hat{\beta}_k Y_t^* - b(Z_t'\hat{\beta}_k) \\ \approx\ & Z_t'\hat{\beta}_k Y_t^* - Z_t'\hat{\beta}_k\mu(Z_t'\beta_k) + Z_t'\beta_k\mu(Z_t'\beta_k) - b(Z_t'\beta_k) + 1/2(\hat{\beta}_k - \beta_k)'H_t(\beta_k)(\hat{\beta}_k - \beta_k) \end{aligned}$$

If the model is specified correctly, that is, if the true model is $Z_t'\beta_k$, then the first two terms should be cancelled out as $E_{Y_t^*|S_{t,k}^*=1}(Y_t^*) = Z_t'\hat{\beta}_k\mu(Z_t'\beta_k)$.

Lemma 1

If $[var(\hat{\beta})]^{-1}$ exists and the conditions for asymptotic normality hold, which are
(i) link function $g(u)$ is twice continuously differentiable and $det(\partial u/\partial\gamma) \neq 0$;
(ii) $\lambda_{min}\mathbf{f}_n \to \infty$, where $\mathbf{f}_n$ is the Fisher information and λ is the smallest eigenvalue;
then for GLM with canonical link $g(\mu_t) = \gamma_{x_t} = Z_t'BS_t$,

$$\begin{aligned} & E_{Y^*,S^*}\left[\log[f(Y^*|S^*;\hat{\psi})] - c(Y^*,\hat{\tau}^2)\right] \\ =\ & E_{Y^*,S^*}\sum_{t=1}^{T}\sum_{k=1}^{K} S_{t,k}^*\left[\frac{Z_t'\hat{\beta}_k Y_t^* - b(Z_t'\hat{\beta}_k)}{\hat{\tau}^2}\right] \\ \approx\ & \sum_{t=1}^{T}\sum_{k=1}^{K} P(S_{t,k}^*=1)\left[\frac{Z_t'\beta_k\mu(Z_t'\beta_k) - b(Z_t'\beta_k)}{\hat{\tau}^2} + \frac{1/2(\hat{\beta}_k - \beta_k)'H(\beta_k)(\hat{\beta}_k - \beta_k)}{\hat{\tau}^2}\right] \end{aligned} \tag{2.23}$$

Lemma 1 handles the expectation with respect to Y^*, thus provides a convenient step for the following approximation. The rest is to approximate the expectation on Y.

Lemma 2

Suppose τ^2 is known and the following conditions hold
(i) the hidden Markov chain is stationary and the transition matrix is ergodic;

(ii) the expectation of the second derivatives of the conditional distribution (exponential family) density with respect to the parameters exists;
(iii) there exists a $\delta > 0$ such that $P(\rho_0(Y_1) = \infty | X_1 = i) < 1$ for all a, where

$$\rho_0(y) = sup_{|\theta-\theta_0|<\delta} max_{1\le i,j\le K} \frac{f(y|x=i)}{f(y|x=j)};$$

then

$$E_Y E_{Y^*,S^*} \log f(Y^*; \hat{\beta}) \approx \log f(\mu; \beta) - DK/2.$$

where D is the dimension of β_k.

Proof. We start with GLM then extend to HMM-GLM. Some of the existing results for GLM can be used, such as $\hat{\beta}$ is asymptotic Normal with $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \tau^2(Z'G^{-1}Z)^{-1}$, where $G = diag(g_\mu(Z_1'\beta), \ldots, g_\mu(Z_T'\beta))$ (McCulloch, 2001). In addition, under canonical link $vg_\mu = 1$. Assume $[Var(\hat{\beta}_k)]^{-1}$ exists, we have

$$\sum_{t=1}^{T} H_t(\beta_k) = \sum_{t=1}^{T} Z_t v(Z_t'\beta_k) Z_t' = Z'V_kZ = Z'G_k^{-1}Z \approx \tau^2[Var(\hat{\beta}_k)]^{-1}$$

where $V_k = diag(v(Z_1'\beta_k), \ldots, v(Z_T'\beta_k))$. Thus,

$$\sum_{t=1}^{T} \frac{(\hat{\beta}_k - \beta_k)' H_t(\beta_k)(\hat{\beta}_k - \beta_k)}{\tau^2} \simeq \chi^2_D.$$

Hence, for GLM

$$E_Y \left[\sum_{t=1}^{T} \frac{1/2(\hat{\beta}_k - \beta_k)' H_t(\beta_k)(\hat{\beta}_k - \beta_k)}{\tau^2} \right] \approx D/2.$$

Similarly, in HMM-GLM setting, for any $k \le K$,

$$\begin{aligned} -E\left[\frac{\partial^2 logL(\theta; Y^*, S^*)}{\partial\beta_k\partial\beta_k'}\right] &= -\frac{\partial^2 \sum_{t=1}^{T} P(S_{t,k}^* = 1)[Z_t'\beta_k Y_t^* - b(Z_t'\beta_k)]}{\partial\beta_k\partial\beta_k'}\Big|_{\beta_k=\hat{\beta}_k} \\ &= \frac{\partial^2 \sum_{t=1}^{T} P(S_{t,k}^* = 1) b(Z_t'\beta_k)}{\partial\beta_k\partial\beta_k'}\Big|_{\beta_k=\hat{\beta}_k} \end{aligned}$$

and $\sum_{t=1}^{T} P(S_{t,k}^* = 1) H_t(\beta_k) = Z' G_k^{-1} W_k Z$, where W_k is the same as defined in the previous section. Thus $Var(\hat{\beta}_k) \approx \tau^2 (Z' G_k^{-1} W_k Z)^{-1}$. With the asymptotic normality of $\hat{\beta}_k$, the following result holds

$$E_Y \left[\sum_{t=1}^{T} P(S_{t,k}^* = 1) \frac{1/2(\hat{\beta}_k - \beta_k)' H_t(\beta_k)(\hat{\beta}_k - \beta_k)}{\tau^2} \right] \approx D/2. \tag{2.24}$$

The conclusion in Lemma 2 indeed coincides with AIC, which again verifies that our initiatives, which is to approximate Kullback-Leibler divergence, is the same and our work is on the right track. To seek improvement, we think further for the situation when τ^2 is unknown, which is usually the case. This approximation is more difficult as now we need to work on

$$E_Y \left[\frac{\sum_{t=1}^{T} P(S_{t,k}^* = 1) 1/2(\hat{\beta}_k - \beta_k)' H_t(\beta_k)(\hat{\beta}_k - \beta_k)}{\hat{\tau}^2} \right]. \tag{2.25}$$

We have shown in Lemma 2 that the numerator in equation (2.25) is asymptotically χ^2 distribution. If the denominator is also a χ^2 distribution and is independent of the numerator, then the expectation of a F statistics may be the solution to this problem.

Recall that in GLM, the nuisance parameter is estimated by

$$\hat{\tau}^2_{GLM} = \frac{1}{N - D} \sum_{n=1}^{N} \frac{(y_n - \hat{\mu}_n)^2}{v(\hat{\mu}_n)}$$

The sum of squares of standardized residuals $\sum (y - \hat{\mu})^2 / v(\hat{\mu})$ is a generalized Pearson χ^2 statistics (McCullagh's book 1983, chapter 2). Modify it to HMM-GLM as we mentioned in the previous section, we have an estimator for τ_k^2 given hidden state k as

$$\hat{\tau}_k^2 = \frac{1}{c_1} \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{(y_t - \hat{\mu}_{t,k})^2}{v(\hat{\mu}_{t,k})} \hat{\xi}_{t,k}$$

where $c_1 = T - D$. Our interest is to find the degree of freedom for it.

Consider in GLM

$$\sum_{t=1}^{T} \frac{(Y_t - \hat{\mu}_t)^2}{v(\hat{\mu}_t)} = \sum_{t=1}^{T} \frac{[Y_t - \mu(Z_t'\beta)]^2 - [\mu(Z_t'\beta) - \mu(Z_t'\hat{\beta})]^2 + 2[Y_t - \hat{\mu}_t][\mu(Z_t'\beta) - \mu(Z_t'\hat{\beta})]}{v(\hat{\mu}_t)}$$

After taking expectation on Y, the first term in numerator becomes $Var(Y_t)$ which equals to $\tau^2 v_t$. The expectation on the second term in the numerator is $Var(\hat{\mu}_t)$. By the asymptotic properties of $\hat{\beta}$, Delta method and the property of canonical link $vg_\mu = 1$, $\mu(Z_t'\hat{\beta})$ is asymptotic Normal with mean $\mu(Z_t'\beta)$ and variance $v_t^2\tau^2 Z_t'(Z'G^{-1}Z)^{-1}Z_t$. Thus,

$$\begin{aligned}
& \sum_{t=1}^{T} E[\mu(Z_t'\beta) - \mu(Z_t'\hat{\beta})]^2/v_t \\
\approx & \sum_{t=1}^{T} tr[v_t\tau^2 Z_t'(Z'G^{-1}Z)^{-1}Z_t] \\
= & \sum_{t=1}^{T} tr[\tau^2 Z_t v_t Z_t'(Z'G^{-1}Z)^{-1}] \\
= & \tau^2 tr[\sum_{t=1}^{T} Z_t v_t Z_t'(Z'G^{-1}Z)^{-1}] = D\tau^2
\end{aligned}$$

The third term in the numerator would vanish after taking expectation since $E(\hat{\mu}_t) = \mu_t$, $c_T \to 0$ as $T \to \infty$. Hence,

$$E_Y \sum_{t=1}^{T} \frac{(Y_t - \hat{\mu}_t)^2}{v(\hat{\mu}_t)} \approx \tau^2(T - D). \tag{2.26}$$

and $\chi^2/\tau^2 \simeq \chi^2_{T-D}$.

Assumed independence between the denominator and numerator, based on the previous results we have

$$E_Y \left[\frac{\sum_{t=1}^{T} (\hat{\beta}_k - \beta_k)' H_t(\beta_k)(\hat{\beta}_k - \beta_k)}{\hat{\tau}^2} \right] \tag{2.27}$$

$$\approx E[\frac{Dc_1}{T-D} \frac{\chi^2_D/D}{\chi^2_{(T-D)}/(T-D)}] \approx \frac{Dc_1}{T-D-2}. \tag{2.28}$$

No matter c_1 equals to $T - D$ or $T + D$, as $T \gg D$, this result reduces to AIC. This supports the independence assumption.

To generalize the above results to HMM-GLM setting, we need to construct an estimator of τ^2 such that it tailors HMM-GLM and gives a better criterion.

Let

$$\hat{\tau}^2 = \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_k^2 = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{c_1} \sum_{t=1}^{T} \frac{(y_t - \hat{\mu}_{k,t})^2}{v_{k,t}} I(s_{k,t} = 1) \tag{2.29}$$

Similarly, taking expectation on $\hat{\tau}^2$ gives

$$\begin{aligned}
E[\hat{\tau}^2] &= \frac{1}{K}\frac{1}{c_1}\sum_{k=1}^{K}\sum_{t=1}^{T} E\left[\frac{(y_t-\hat{\mu}_{k,t})^2}{v_{k,t}}|S_{k,t}\right] E(S_{k,t}) \\
&\approx \frac{1}{c_1 K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[E[\frac{(y_t-\mu_{k,t})^2}{v_{k,t}}|S_{k,t}]E(S_{k,t}) - E[\frac{(\mu_{k,t}-\hat{\mu}_{k,t})^2}{v_{k,t}}|S_{k,t}]E(S_{k,t})\right] \\
&= \frac{1}{c_1 K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\tau^2 - E[\frac{(\hat{\mu}_{k,t}-\mu_{k,t})^2}{v_{k,t}}|S_{k,t}])\right] E(S_{k,t}) \\
&\approx \frac{1}{c_1 K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\tau^2 - v_{k,t}\tau^2 Z_t'(Z'G_k^{-1}W_k Z)^{-1}Z_t\right] E(S_{k,t}) \\
&= \frac{\tau^2}{c_1 K}[T-KD] \qquad (2.30)
\end{aligned}$$

Since $\sum_{k=1}^{K} P(S_{k,t}=1)=1$, the first term goes to $T\tau^2$.

Thus, for HMM-GLM, under the same condition as stated in Lemma 2, $c_1 K\hat{\tau}^2/\tau^2$ approximates χ^2 distribution,

$$c_1 K E_Y \hat{\tau}^2 \approx \tau^2(T-KD). \qquad (2.31)$$

and

$$\begin{aligned}
& E_Y\left[\sum_{k=1}^{K}\sum_{t=1}^{T} P(S_{k,t}^*=1)\frac{1/2(\hat{\beta}_k-\beta_k)'H_t(\beta_k)(\hat{\beta}_k-\beta_k)}{\hat{\tau}^2}\right] \\
\approx\ & 1/2E[\frac{K^2 D c_1}{T-KD}\frac{\chi^2_{KD}/KD}{\chi^2_{(T-KD)}/(T-KD)}] \\
\approx\ & \frac{1}{2}\frac{K^2 D c_1}{T-KD-2}.
\end{aligned}$$

Hence, the proposed criterion for the HMM-GLM is

$$C_{hmm-glm} = \log L(\hat{\theta};Y) - \frac{1}{2}\frac{K^2 D c_1}{T-KD-2}. \qquad (2.32)$$

2.5 SIMULATION

In this section, we conduct a simulation study to compare the proposed model selection criterion with AIC and BIC. We consider large, medium and small sample sizes. Here the sample size refers to the number of observation time points in a sequence in our model ($T = 50, 100, 250$). We generate data from two typical distributions in the exponential family: Gaussian and Poisson. For each setting, we repeat the simulation 200 times. The comparison is based on the number of times that the criteria choose the right number of variables and the right number of hidden states. The simulation is implemented using R package 'RHmm'.

The simulation setting for the Gaussian case is a HMM with three hidden states and state-specific coefficient vectors $\beta_1 = (1, 2, 3)$, $\beta_2 = (4, 3, 2)$, $\beta_3 = (-1, -2, -3)$ for two predictors. The scaling parameter σ^2 is set to 1 for all three hidden states. The transition matrix is

$$A_1 = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$$

The simulation setting for the Poisson case is a HMM with two hidden states and state-specific coefficient vectors $\beta_1 = (0.1, -0.8, -0.5, -0.4)$, $\beta_2 = (0.8, 0.5, 0.7, 0.3)$ for 3 predictors. The transition matrix is

$$A_2 = \begin{bmatrix} 0.65 & 0.35 \\ 0.25 & 0.75 \end{bmatrix}$$

Tables 1, 2, and 3 show the rate of choosing the right K and D jointly and separately. From the results, we see that the proposed criterion works better than AIC and BIC for small and medium sample sizes. When there are sufficient number of observations, BIC performers better than the other two.

Table 1: Model selection criterion comparison (% of times correct in K and D)

T	K	D	Distribution	AIC	BIC	$C_{hmm-glm}$
50	3	2	Gaussian	0.15	0.15	0.46
100	3	2	Gaussian	0.32	0.49	0.67
250	3	2	Gaussian	0.46	0.89	0.66
50	2	3	Poisson	0.62	0.42	0.58
100	2	3	Poisson	0.74	0.74	0.81
250	2	3	Poisson	0.79	0.98	0.89

Table 2: Model selection criterion comparison (% of times correct in K)

T	K	D	Distribution	AIC	BIC	$C_{hmm-glm}$
50	3	2	Gaussian	0.29	0.13	0.58
100	3	2	Gaussian	0.53	0.35	0.76
250	3	2	Gaussian	0.52	0.95	0.71
50	2	3	Poisson	0.97	1.00	0.99
100	2	3	Poisson	0.96	1.00	0.99
250	2	3	Poisson	0.93	1.00	0.98

Table 3: Model selection criterion comparison (% of times correct in D)

T	K	D	Distribution	AIC	BIC	$C_{hmm-glm}$
50	3	2	Gaussian	0.40	0.83	0.87
100	3	2	Gaussian	0.58	0.92	0.90
250	3	2	Gaussian	0.67	0.97	0.92
50	2	3	Poisson	0.62	0.42	0.55
100	2	3	Poisson	0.76	0.74	0.82
250	2	3	Poisson	0.85	0.98	0.91

2.6 DISCUSSION

In this section, we first established a general form of estimation procedures for HMM-GLM, focusing on how to obtain the coefficients of the covariates. Then we proposed a model selection criterion by approximating the Kullback-Leibler divergence. Our major contribution is to approximate the emission distribution. The derived criterion is a function of both K and D, thus serves the purpose for joint selection. The conditions in the theoretic derivation part are based on previous work on asymptotic normality of estimators. We made certain assumptions without proof in the process to obtain the proposed criterion; therefore we conducted a simulation study to check that they are reasonable. We compared our criterion with AIC to make sure no large difference between the two because both of them are approximating the Kullback-Leibler divergence. In the simulation, we find the proposed criterion performs better in small and medium sample size cases, which is more difficult. For a better approximation, it is necessary to consider approximating the non-GLM part of the likelihood functions.

3.0 CLUSTERING SEQUENCES WITH HIDDEN MARKOV MODELS

3.1 INTRODUCTION OF HIDDEN MARKOV CLUSTERING

Clustering is a data analysis technique to explore partitions of a collection of cases according to certain criteria in order to find less heterogeneous groups. It is important and widely used in modern statistical data analysis. Existing clustering methods can be summarized into three types: partition clustering, hierarchical clustering, and mixture model clustering.

Sequence clustering is a special kind of clustering problem where the individual cases are sequences of observations that are temporally correlated. This dependence structure raises more challenge such that many existing clustering methods become unsuitable, because they cannot take into account the correlations between observations at different time points. Yet motivated by an increasing number of applications, sequence clustering has attracted considerable interest in recent years. It is especially important in fields like bioinformatics and the study of the Internet users behavior analysis. In bioinformatics, sequence clustering algorithms are used to group biological sequences, such as protein and DNA. In studies of the Internet, people use sequence clustering methods on the click-path data to explore different patterns that users navigate or browse a website (Cadez, 2003).

HMMs have the advantage of modeling state changes in sequential data, thus sequence clustering with HMMs is of interest. It was first studied by Smyth (1997), who used the log-likelihood to measure the discrepancy between two sequences and then applied hierarchical clustering on the resulting distance matrix. Later, others considered various metrics to compute the distance or similarity between HMM sequences. Panuccio et al. (2002) developed a model-based method and Bicego et al. (2003) proposed a similarity-based method. Yet these methods do not scale well for large data problems. The main reason is that con-

structing the distance matrix based on the pairwise likelihood of HMMs is computational expensive. Another way to consider sequence clustering with HMMs is using a parametric method. Coviello et al. (2012) proposed a variational hierarchical EM algorithm for clustering HMMs, using the parameters to characterize the cluster centers. This method assigns points to cluster centers by Kullback-Leibler divergence.

The aim of our study is to develop an algorithm that is less computationally expensive and easier to interpret for the HMM sequence clustering problem. In HMMs, the observed sequences are governed by the emission probabilities and transition matrices. In other words, when the sequences can be well represented by their emission probabilities and transition matrices, these two features can be used to identify the observed sequence types. For sequences with hundreds or more time points, the emission and transition features can be viewed as a lower-dimension representation of the original data. Garcia et al. (2011) proposed a sequence clustering method with HMM setting based on the transition matrix induced in a common HMM. This method differs from previous methods in that it avoids computing the likelihood distance matrix. However, clustering HMM sequences only relies on the transition matrices while ignoring the possible discrepancy in the emission distribution, may then weaken the ability to distinguish sequences. A better way we considered is based on both the emission and transition features of the HMM sequences. However, to implement the idea, one challenge lies in finding a suitable scalar to measure the discrepancy on the emission and transition features. Another challenge is combining the discrepancy of emission probability and discrepancy of transition patterns into a synthetic distance measure. In our proposed method, we use the Bhattacharyya affinity for both transition and emission distribution. The Bhattacharyya distance is often used to measure the similarity of two discrete or continuous probability distributions, and is a special case of probability product kernels. Given any two discrete distributions p and q, the Bhattacharyya affinity is defined as

$$B(p, q) = \sum_x \sqrt{p(x)q(x)}; \tag{3.1}$$

it is related to Hellinger distance: $B(p, q) = 1 - H^2(p, q)$. Thracker et al. (1997) explains that the Bhattacharyya affinity is preferred over chi-square statistics in large distance matrices

because the chi-square statistics are poor at handling empty cells but the Bhattacharyya affinity doesn't have this problem.

Jebara et.al. (2007) have laid down the foundations for the use of the Bhattacharyya affinity to measure the similarity between two HMM sequences. They developed a semi-parametric method which combines HMM settings with spectral clustering. In addition, their method takes into account the hidden state matching-comparison problem by including affinity measures of all permutations of the states. Our work is based on this method to seek improvement with a more efficient algorithm. We show that the iterative computation of the affinity measure can be approximated by a function of the emission distribution and transition matrix. Our main contribution is in finding an efficient way to obtain the affinity matrix. After obtaining the affinity matrix, we apply spectral clustering method to output the final clustering results.

Spectral clustering is a very popular clustering algorithm in modern data mining. Besides the successes in both empirical studies and synthetic data, it has the appealing advantage that is easy to implement. Applied it to HMM sequences, no more assumptions or further complicated model structures are imposed, other than the HMM setting. In addition, it generally has better performance compared to some traditional clustering algorithms, such as K-means. The name "spectral" comes from the fact that the method uses eigenvectors of the affinity matrix. Instead of directly applying the K-means algorithm to the affinity matrix, spectral clustering applies K-means to the derived eigenvector matrix of the affinity matrix. The intuition is somewhat similar to principal component analysis (PCA), which maps the original data matrix to a lower dimensional space spanned by the largest eigenvectors of the data matrix itself. K-means is then used after mapping the affinity matrix.

Both simulations and real data comparison show that our method is competitive and it improves efficiency. Although the methodology should be applicable to general emission distributions, in this study we focus on the exponential family cases since their Bhattacharyya affinity formulas are in closed form.

The rest of this section is organized as follows. In section 3.2, we describe how we developed the algorithm. In section 3.3, we conduct a simulation study to test the performance of our proposed method. In section 3.4, there is a real data comparison of our proposed

method and Jebara's method (2007). We denote their method as PPK hereafter. In section 3.5, we conclude with a discussion of the proposed method.

3.2 METHOD DEVELOPMENT

For most of this section, we continue to use the same notation as in the previous chapter. Let A be the transition matrix, π be the initial distribution, θ be the emission parameters, and K be the number of hidden states. Denote observed chain by Y and hidden chain by X. Again, we assume the hidden Markov chain is homogeneous. We begin with the forward-backward algorithm, which is used to compute the likelihood of the observed sequence. Denote the likelihood function for an observed sequence $\mathbf{Y_n}$ by $L(\theta, A, \pi; \mathbf{Y_n})$. Let α_t be the 'forward' vector and $\mathbf{f}_t$ be a diagonal matrix,

$$\alpha_t(k) = P(Y_1, \ldots, Y_t, X_t = k) \tag{3.2}$$

$$\mathbf{f}_t = diag[f(y_t|x_t = 1), \ldots, f(y_t|x_t = K)]. \tag{3.3}$$

For any sequence with T_n number of observations, the forward-backward algorithm computes the observed likelihood:

$$\alpha_1 = \pi \mathbf{f}_1 \tag{3.4}$$

$$\alpha_{t+1} = \alpha_t A \mathbf{f}_{t+1}$$

$$L(\theta; Y_1, \ldots, Y_{T_n}) = \sum_{k=1}^{K} \alpha_{T_n}(k) = \sum_{k=1}^{K} \pi [\prod_{t=1}^{T-1} \mathbf{f}_t A] \mathbf{f}_T.$$

.

Note that for any t, adjusted with the normalized constant $tr(\mathbf{f}_t)$, $\mathbf{f}_t A$ becomes a probability square matrix. The Bhattacharyya affinity for two sets of parameters $(A^{(1)}, \theta^{(1)})$ and $(A^{(2)}, \theta^{(2)})$ is

$$\begin{aligned} & \int \sum_{i=1}^{K} \sum_{j=1}^{K} \sqrt{f_i(y_t)^{(1)} a_{ij}^{(1)} f_i(y_t)^{(2)} a_{ij}^{(2)}} dy_t \qquad (3.5) \\ = & \sum_{i=1}^{K} \sum_{j=1}^{K} B_{f_i} \sqrt{a_{ij}^{(1)} a_{ij}^{(2)}} \\ = & B_f \circ B_A \end{aligned}$$

where B_{f_i} is the Bhattacharyya affinity of the emission distributions at state i, and $B_f = (B_{f_1}, \ldots, B_{f_K})$; B_{A_i} is the Bhattacharyya affinity of row i in the transition matrices, and $B_A = (B_{A_1}, \ldots, B_{A_K})$; $B_f \circ B_A$ is the inner product of the two vectors. Since we integrate out y_t, that term is constant for different time points. This gives us the great advantage that we may avoid iterative computation to obtain the affinity for all time points. With this formula, we are getting close to seeing how to represent the likelihood with the transition and emission distributions.

Next, instead of looking at the Bhattacharyya affinity for the likelihood $L(\theta, A, \pi; Y_1, \ldots, Y_T)$, it is easier to see the importance of the emission and transition features if we look at the Bhattacharyya affinity for α_T. Without loss of generosity, we assume that the number of observation time points of the two sequences are the same, then the Bhatacharyya affinity for $\alpha_T^{(1)}$ and $\alpha_T^{(2)}$ is

$$\int L(\theta^{(1)}; Y_1, \ldots, Y_T, X_T) L(\theta^{(2)}; Y_1, \ldots, Y_T, X_T) dy \propto [B_f \circ B_A]^{T-1}. \tag{3.6}$$

Note that α_T is the last step in the forward-backward algorithm to obtain the likelihood; it is proportional to the likelihood of the observed sequence. It is a function of the transition matrix and emission distribution. Hence, we can use the value of $[B_f \circ B_A]^{T-1}$ to measure the pairwise affinity of HMM sequences. Denote it as $\mathbf{B}_{hmm}$ hereafter. With a little extra effort, we take into account the state permutation problem, by considering both the sum of the permuted affinity measures and the maximum of the permuted affinity measures.

We now illustrate our method using the the multivariate normal distribution. Suppose that there are two HMM sequences with parameters $(\mu^{(1)}, \Sigma^{(1)}, A^{(1)})$ and $(\mu^{(2)}, \Sigma^{(2)}, A^{(2)})$, respectively. Assume that there are two hidden states. For each state i

$$B_{f_i} = \frac{|2U_i|^{1/2}}{\sqrt{|\Sigma_i^{(1)}|^{1/2} |\Sigma_i^{(2)}|^{1/2}}} \exp[-\frac{1}{4}(\mu_i^{(1)} - \mu_i^{(2)})' M_i^{-1} (\mu_i^{(1)} - \mu_i^{(2)})] \tag{3.7}$$

where

$$M_i = \Sigma_i^{(1)} + \Sigma_i^{(2)} \quad \text{and} \quad U_i^{-1} = [\Sigma_i^{(1)}]^{-1} + [\Sigma_i^{(2)}]^{-1}.$$

Then

$$\mathbf{B}_{hmm} = \left[B_{f_1} \left(\sqrt{a_{11}^{(1)} a_{11}^{(2)}} + \sqrt{a_{12}^{(1)} a_{12}^{(2)}} \right) + B_{f_2} \left(\sqrt{a_{21}^{(1)} a_{21}^{(2)}} + \sqrt{a_{22}^{(1)} a_{22}^{(2)}} \right) \right]^{T-1}. \tag{3.8}$$

So far, we have developed a non-iterative way to obtain the affinity matrix. The remaining task is to apply spectral clustering methods to obtain the final clustering assignments. The steps of our proposed HMM sequences clustering method are summarized below.

(1) Fit an HMM to each sequence and obtain the transition matrix and emission parameter estimates. For N sequences, there are a list of N transition matrices and N sets of emission parameters.

(2) For each pair of sequences, compute $\mathbf{B}_{hmm}$ using the corresponding transition matrices and emission parameters. The affinity matrix consists of $\frac{N(N-1)}{2}$ pairs of $\mathbf{B}_{hmm}$. Denote the affinity matrix as $\mathbf{B}_{hmm}$.

(3) Apply spectral clustering on the obtained affinity matrix. Let $L_B = D_B^{-1/2} \mathbf{B}_{hmm} D_B^{-1/2}$, where D_B is the diagonal matrix with non-zero element on row i as the row sum of $\mathbf{B}_{hmm}$. Suppose we want to have l clusters; then we put l eigenvectors corresponding to the l largest eigenvalues of L_B into the matrix $V_B = [v_1, \ldots, v_l]$. Next, normalize V_B so that each row has unit length. Finally, apply K-means algorithms to this spectral matrix V_B.

3.3 SIMULATION

In the first simulation, we repeat the setting in Smyth's (1997) paper. The model is an HMM with two hidden states and a univariate observed variable. The emission distributions are $N(0, 1)$ and $N(3, 1)$ for the two states respectively. There are two clusters which have different transition matrices A_1 andA_2. Each simulated dataset contains 20 sequences from each cluster. The length of all sequences are set to be $T = 200$ as in Smyth's paper. For further comparisons, we try to decrease the number of time points to test our method's performance. Each simulation setting is repeated 200 times.

$$A_1 = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}$$

In 200 runs, the average correctness of our proposed method is 99.65% with standard deviation (sd) 0.01. When $T = 150$, the average correctness is 98.75% sd 0.01. For $T = 100$, the average correctness maintains 99.50% with sd less than 0.01. When $T = 50$, the average correctness is 82.78% sd 0.02.

In addition, we find that when T is large enough (e.g. $T \geq 150$ in this simulation), using $\log(T-1)$ as the power term in $\mathbf{B}_{hmm}$ performs nearly as well as using T. This can further reduce the computational time. For smaller T (e.g. $T = 100$), using $\log(T-1)$ degrades the accuracy and is not comparable to $\mathbf{B}_{hmm}$ with power term T.

In the second simulation, we try another situation in which the two clusters differ at the emission distribution only. Let one cluster have emission distribution as $N(0,1)$ and $N(3,1)$ for the two hidden states; another cluster have $N(1,0.5)$ and $N(2,1)$. The transition matrix is the same (A_1) for both clusters.

The result shows that the average correctness of our proposed method is 99.15% sd 0.02 when $T = 200$. When $T = 150$, the average correctness is 98.75% sd 0.01; When $T = 100$, the average corrrectness is 95.9% sd 0.08; when $T = 50$, 91.09% sd 0.12.

3.4 EMPIRICAL STUDY: AUSTRALIAN SIGN LANGUAGE

In addition to our simulation study, we compare our proposed method to PPK (Jebara 2007) with a real data set. The Australian dataset consists of several sign-language gestures (see https://archive.ics.uci.edu/ml/datasets/). Each gesture has 27 instances with an average of 60 'time points'. There are 22 variables recorded for each instance. This dataset has been used in both Jebara's (2007) and Garcia's (2011) study. In our study, we take the first two principal components of the 22 variables and fit them with two-state HMM. Fig.1and Fig 2 show the sequences of 5 pairs of gestures that we use to compare the clustering performance.

Table 4: Clustering algorithms comparison on Australian sign language data (correct percentage)

Method	hot-cold	spend-cost	eat-drink	happy-sad	yes-no
$\mathbf{B}_{hmm}$	100	98	87	83	67
PPK	100	80	93	87	59

The results show that our method is more efficient since we save the iteration cost, while maintaining similar performance (accuracy) as PPK.

3.5 DISCUSSION

In this study, we developed an efficient sequence clustering method with HMMs using representatives of the emission and transition distributions of the HMM sequences. Both the simulation study and an empirical study show that the proposed method is competitive with earlier more computationally intensive methods. However, we acknowledge that to theoretically verify the conditions when our method is suitable, a more detailed study is needed. Also, it would be worth more effort to compare other affinity measures. There are a large number of applications of mobility measures discussed in the social science literature, where the states are taken to be social classes or occupational groups, and in economic studies for credit migration. When constructing a mobility index for transition matrices, some desirable properties are mentioned, such as normalization, monotonicity, immobility and perfect mobility, but Shorrocks (1978) pointed out that it would be impossible to satisfy simultaneously all these properties — for example, normalization, monotonicity and perfect mobility are incompatible. In fact, the χ^2 test statistic was used by Hilton (1971) as an alternative approach to define distance between transition matrices. However, it is not suitable for

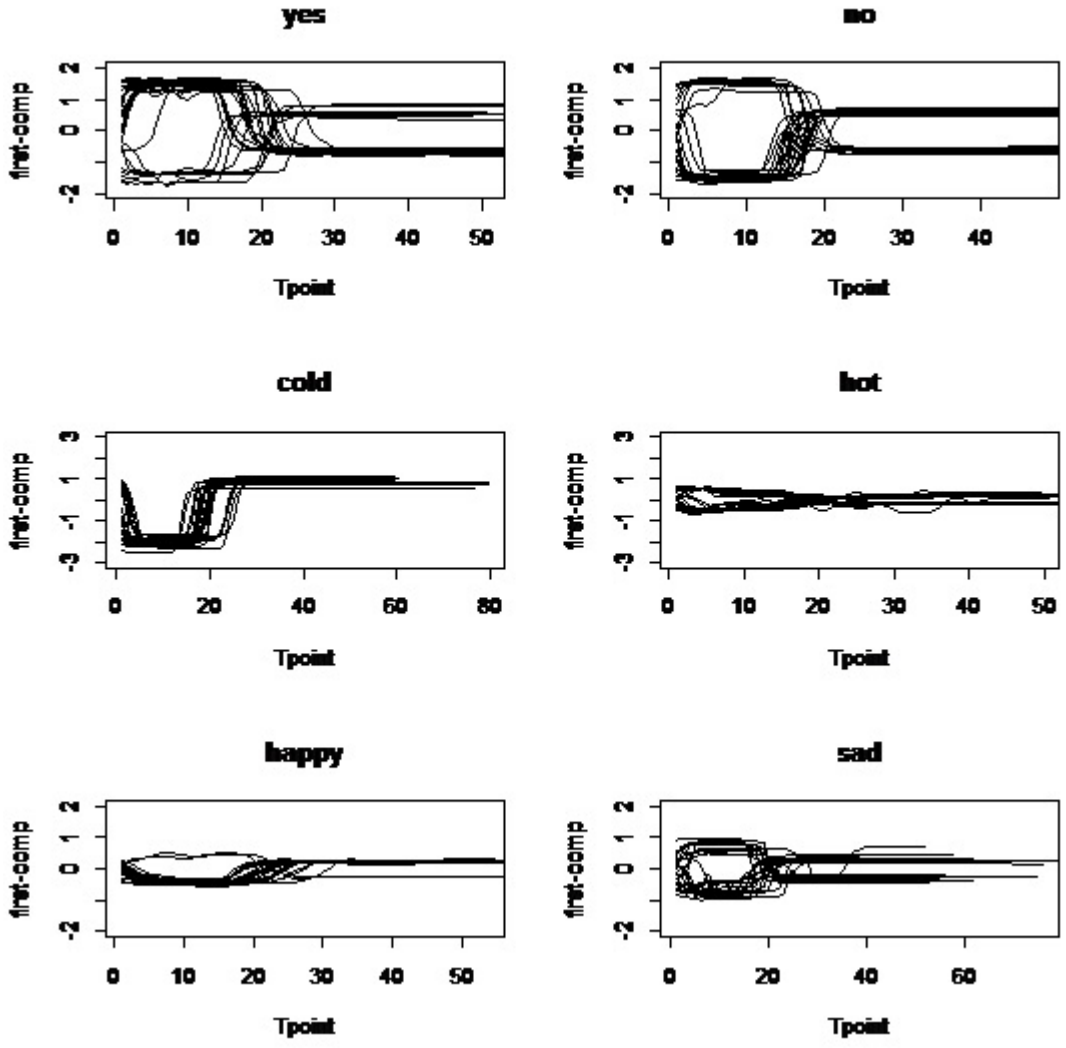

Figure 1: Australian sign language data sample 1

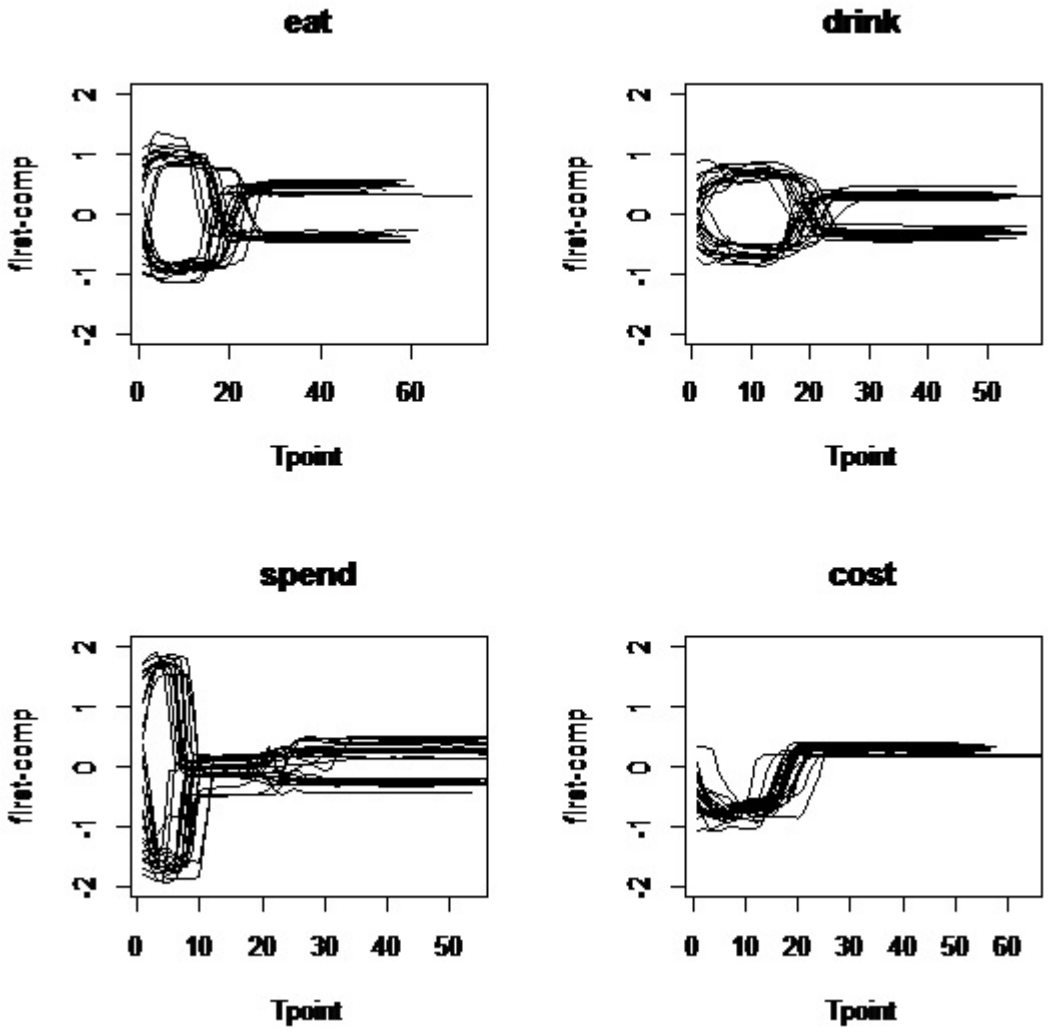

Figure 2: Australian sign language data sample 2

comparing emission probabilities and thus cannot serve our purpose.

Another consideration for people to decide whether sequence clustering with HMMs is a suitable method to test the model assumptions, specifically the Markovian property and its order. Sahalia et al. (2010) proposed a nonparametric test method for the Markov property, based on the Chapman-Kolmogorov equation. Chen, et al. (2012) pointed out that the Chapman-Kolmogorov equation was only a necessary condition for the Markov property. They provided a conditional characteristic function-characterization for the Markov property and use it to construct a test for the Markov property, applying a nonparametric regression method to estimate the conditional characteristic function. The challenge that arises in the HMM is that we cannot directly apply these procedures because of two reasons. First, the above procedures are designed for the observed stochastic process. But according to the model structure of the HMM, we cannot infer that the hidden stochastic process possesses the Markov property from the corresponding observed Markov process. So we cannot use the above procedures to test the Markov property of the observed sequence. Second, we cannot apply these procedures to the estimated hidden state sequence either, because the hidden state sequence is obtained using the Markov assumption. We suggest the use of the existing goodness-of-fit test and model diagnosis methods for checking the Markov assumption. For example, the pseudo-residual model checking procedures provides a way to check the Markov property. If a histogram or quantile-quantile (qq) plot of the uniform pseudo-residuals supports the conclusion that they are U(0,1), that would suggest that the Markov assumption in the model is valid.

For future work, we are interested in trying another approach to implement the clustering based on the transition pattern and emission distribution features of the sequences and comparing the two approaches. The alternative mean consider constructing the similarity matrix using random samples to compute the Hellinger distance. We also expect it would own strength in efficiency.

4.0 CLUSTERING TRAJECTORIES WITH FIRST-ORDER MARKOV MIXTURE MODELS

4.1 INTRODUCTION OF MARKOV MIXTURE MODELS AND BIPOLAR DISORDER STUDY

Bipolar disorder (BD) is a mental illness characterized by recurrent mood episodes ranging from periods of depression or irritability to periods of extreme happiness. Depressive episodes usually include feelings of sadness, lack of enjoyment, low energy, and problems with sleep and appetite. Conversely, manic or hypomanic episodes are periods of extreme happiness that are usually accompanied by high levels of energy, a decreased need for sleep, racing thoughts, and grandiosity. Measuring and making sense of these mood fluctuations over time is of particular interest, but few statistical models have been developed for this to date. Previous studies established a uniform scoring system to quantify the various mood states based on a simulation study (Rao et al., 2006), and applied logistic regression to examine relations between rapid mood changes and other factors such as early age of onset and anxiety (Nwulia et al., 2008). However, these studies did not model mood fluctuations as a stochastic process and the occurrence of a mood switch was based on the individual answering yes to the single question "Have you ever switched back and forth quickly between feeling high and feeling normal or depressed ?"

The purpose of this section is to develop a stochastic model for the mood transitions of BD patients and see whether there exist subgroups of mood fluctuation patterns. Based on clinical relevance, we have focused on four mood states: well (formally known as euthymia), mania/hypomania, depression and mixed (a combination of symptoms of mania/hypomania and depression). Specifically, we are interested in the frequency and patterns of changes

among them. Thus, models of interest are those that can cluster sequences according to the state transition patterns.

For existing clustering methods, such as partitional clustering (e.g. K-means), hierarchical clustering, and spectral clustering, the main challenge in applying them to longitudinal data lies in how to define a distance measure between observed sequences, especially when the observed sequences have different lengths. Model-based clustering methods are often used to analyze longitudinal data. However, most existing methods, such as latent growth models and clustering for functional data (James and Sugar, 2003) that were developed for continuous outcomes are not suitable for discrete state sequences. Although a mixture of generalized linear mixed models could handle discrete outcomes, they cannot model the patterns of mood fluctuations needed to analyze the longitudinal course of individuals with BD (Molenberghs et al., 2005; Komarek et al., 2013). In contrast, Hidden Markov models (HMMs) provide an approach for such data. Researchers have proposed various ways for clustering sequences with HMMs, such as building a likelihood distance matrix of the observed sequences with hierarchical clustering (Smyth, 1998) and using probability kernel product for an affinity measure with spectral clustering (Jebara, 2007). Recently, mixture latent Markov models have been applied to cluster the within-day mood change patterns of healthy subjects, focusing on pleasant-unpleasant states (Crayen et al., 2012). However, in our study the number of observed states is small, and their interpretation is clear, so modeling hidden states is not compelling.

In this study we introduce the Markov mixture model with different transition matrices to find homogeneous clusters. The continuous-time Markov models defined on a finite discrete state space have the advantage of capturing state switching patterns over time using transition matrices, which can be viewed as an instance of data reduction: from hundreds or more observation time points to a small matrix. In addition, this model is flexible enough to handle sequences with various lengths since they are encoded by the transition matrices.

To the best of our knowledge, there are few applications of Markov models for clustering in mental health studies. They are mainly used to learn the association between the Markov chains and certain explanatory variables (Sung et al. 2007; Visser et al. 2002). Integrating Markov models and mixture models for clustering problems was first used to find navigation

patterns on web sites (Cadez et al., 2003). In our current study we provide a more detailed model development. In addition, we conduct simulation studies to investigate the model performance both when model assumptions hold and when there are certain violations of them. The simulation results show that the model is fairly stable for both cases.

The rest of this section is organized as follows. In section 4.2, we describe the Markov mixture model, putting some technical details in an Appendix. In section 4.3, we present a simulation study to investigate the model performance. In section 4.4 we give an application to the motivating bipolar disorder study. We conclude with a discussion in section 4.5.

4.2 MODEL DEVELOPMENT

Consider a Markov chain Y on a finite discrete state space $\mathcal{M} = (1, 2, \cdots, M)$. Let $Y_{n,t}$ be the observation for subject n at time t, and let $\mathbf{Y_n} = (\mathbf{Y_{n,1}}, \ldots, \mathbf{Y_{n,T_n}})$ be the Markov chain for subject n. Let K be the number of clusters in the model and C denote the cluster label. Assuming time-homogeneity, the Markov chains in cluster k are governed by the M-by-M transition matrix $A_k = \{a_{k,ij}\}$ and initial probability $\pi_{\mathbf{k}} = (\pi_{\mathbf{k,1}}, \ldots, \pi_{\mathbf{k,M}})$. The first-order Markov mixture model for the sequence $\mathbf{Y_n}$ is:

$$\begin{aligned} P(\mathbf{Y_n}; \theta) &= \sum_{k=1}^{K} P(C_n = k) P(\mathbf{Y_n} | \mathbf{C_n} = \mathbf{k}; \theta_{\mathbf{k}}) \\ &= \sum_{k=1}^{K} w_k \left[\pi_{k, y_{n,1}} \prod_{t=2}^{T_n} P(Y_{n,t} | Y_{n,t-1}; A_k) \right], \end{aligned} \tag{4.1}$$

where $\theta_k = (\pi_k, A_k)$ is the set of parameters for cluster k, and $w_k = P(C_n = k)$ is the weight of component k in the mixture model, subject to $\sum_{k=1}^{K} w_k = 1$. Since, $P(\mathbf{Y_n}, \mathbf{C_n} = \mathbf{k}) = \mathbf{P(Y_n | C_n = k) P(C_n = k)}$, we can write

$$P(\mathbf{Y}_n, C_n) = \prod_{k=1}^{K} \left[P(\mathbf{Y_n} | \mathbf{C_n} = \mathbf{k}) \mathbf{P}(\mathbf{C_n} = \mathbf{k}) \right]^{I(C_n = k)}, \tag{4.2}$$

so that the joint likelihood based on $(\mathbf{Y}, \mathbf{C})$ for all sequences with Lagrange multipliers $(\lambda, \alpha_k, \beta_{k,j})$ to incorporate the constraints on w, π, and the transition matrices A_k is

$$\begin{aligned} & \log L(\mathbf{Y}, \mathbf{C}) \qquad (4.3) \\ = & \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\log(w_k) + \sum_{i=1}^{M} I(Y_{n,1}=i)\log(\pi_{k,i}) + \sum_{i=1}^{M}\sum_{j=1}^{M} b_{n,ij}\log(a_{k,ij})\right] I(C_n = k) \\ + & \lambda\sum_{k=1}^{K}(w_k - 1) + \sum_{k=1}^{K}\alpha_k\sum_{i=1}^{M}(\pi_{k,i}-1) + \sum_{k=1}^{K}\sum_{i=1}^{M}\beta_{k,i}\sum_{j=1}^{M}(a_{k,ij}-1) \end{aligned}$$

We assign the observed sequence $\mathbf{Y_n}$ to the cluster that has the largest posterior probability $P(C_n|\mathbf{Y_n};\eta)$, where $\eta = (\theta_1, \ldots, \theta_K, w_1, \ldots, w_K)$ and

$$P(C_n = k|\mathbf{Y_n};\eta) = \frac{\mathbf{w_k}\mathbf{P}(\mathbf{Y_n}|\mathbf{C_n}=\mathbf{k};\theta_\mathbf{k})}{\sum_{\mathbf{k=1}}^{\mathbf{K}}\mathbf{w_k}\mathbf{P}(\mathbf{Y_n}|\mathbf{C_n}=\mathbf{k};\theta_\mathbf{k}).} \qquad (4.4)$$

We use the EM algorithm for parameter estimation. Let $L(\eta|C_1, \ldots, C_n, \mathbf{Y_1}, \ldots, \mathbf{Y_n})$ be the joint likelihood of the unknown cluster assignments and observed sequences. Because all the observed sequences are independent of each other, the joint likelihood is the product of the individual ones. The E-step is

$$\begin{aligned} Q(\eta, \eta^{old}) &= \mathbf{E_{C|Y}}\left[\log \mathbf{L}(\eta|\mathbf{C_1}, \ldots, \mathbf{C_n}, \mathbf{Y_1}, \ldots, \mathbf{Y_n})|\mathbf{Y_1}, \ldots, \mathbf{Y_n}\right]_{\eta^{\mathbf{old}}} \qquad (4.5) \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N}\log[w_k P(\mathbf{Y_n}|\mathbf{C_n}=\mathbf{k};\theta_\mathbf{k})]\mathbf{P}(\mathbf{C_n}=\mathbf{k}|\mathbf{Y_n};\eta^{\mathbf{old}}) \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\log(w_k) + \sum_{i=1}^{M} I(Y_{n,1}=i)\log(\pi_{k,i}) + \sum_{i=1}^{M}\sum_{j=1}^{M} b_{n,ij}\log(a_{k,ij})\right] P_{nk}^{old}, \end{aligned}$$

where $b_{n,ij}$ is the number of transitions from state i to state j for subject n and $P_{nk}^{old} = P(C_n = k|\mathbf{Y_n};\eta^{\mathbf{old}})$ is the posterior probability of cluster assignment given by (43).

The parameters to be optimized in the M-step are (π_k, A_k, w_k). They are subject to the constrains $\sum_{k=1}^{K} w_k = 1$, $\sum_{i=1}^{M}\pi_{k,i} = 1$ and $\sum_{j=1}^{M} a_{k,ij} = 1$. In the Appendix we show that the updated formulas for each iteration are

$$w_k^{new} = \frac{\sum_{n=1}^{N} P_{nk}^{old}}{\sum_{k=1}^{K}\sum_{n=1}^{N} P_{nk}^{old}}, \qquad (4.6)$$

$$\pi_{k,i}^{new} = \frac{\sum_{n=1}^{N} P_{nk}^{old} I(y_{n,1}=i)}{\sum_{i=1}^{M}\sum_{n=1}^{N} P_{nk}^{old} I(y_{n,1}=i)}, \qquad (4.7)$$

$$a_{k,ij}^{new} = \frac{\sum_{n=1}^{N} P_{nk}^{old} b_{n,ij}}{\sum_{j=1}^{M}\sum_{n=1}^{N} P_{nk}^{old} b_{n,ij}}, \qquad (4.8)$$

where $I(B)$ is the indicator of the set B. Notice that at each step the parameter estimates w^{new}, π^{new}, and $a_{k,ij}^{new}$ satisfy the constraint that they are probability vectors.

4.3 SIMULATION

Previous work (Cadez et al., 2003) did not study the properties of Markov mixture models, either theoretically or by simulation. In this section, we use simulation for both situations when model assumptions hold and when there are violations of the time homogeneity or Markovian assumption.

Motivated by the bipolar study and preliminary results, we consider six clusters on a four-state space. Three of them are homogeneous, one called a "stayer" cluster (TM_s), one called a "mover" cluster (TM_m), and one called a "sub-chain" cluster (TM_{sub}) in which the Markov chains only take values in a subset of the state space. In the first simulation, we mix sequences from these three clusters. The other three clusters violate the model assumptions in terms of homogeneity and Markovian property. Two of them are non-homogeneous, one is a "slow-change non-homogeneity" cluster (TM_{sc}) in which the transition matrix changes from stayer to mover type as time goes by; the other is a "quick-change non-homogeneity" cluster (TM_{qc}) with a change happening in a single step from TM_m to TM_s in the middle of the observation period. Finally we consider a "noisy" cluster (TM_n) which consists of independent sequences without Markov dependence between consecutive observations. An example from each cluster is given in Figure 3. The transition matrices of the simulated clusters are given below.

$$TM_s = \begin{bmatrix} 0.9 & 0.05 & 0.01 & 0.04 \\ 0.1 & 0.85 & 0.02 & 0.03 \\ 0.02 & 0.01 & 0.95 & 0.02 \\ 0.02 & 0.03 & 0.1 & 0.85 \end{bmatrix} \quad TM_m = \begin{bmatrix} 0.6 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.1 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.1 & 0.4 \end{bmatrix}$$

$$TM_{sub} = \begin{bmatrix} 0.8 & 0.1 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.1 & 0.0 \\ 0.1 & 0.1 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \quad TM_n = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$TM_{sc} = \begin{bmatrix} 0.9\cos^2(\frac{t\pi}{u}) & 0.9\sin^2(\frac{t\pi}{u}) & 0.05 & 0.05 \\ 0.1 & 0.9\cos^2(\frac{t\pi}{u}) & 0.3\sin^2(\frac{t\pi}{u}) & 0.6\sin^2(\frac{t\pi}{u}) \\ 0.3\sin^2(\frac{t\pi}{u}) & 0.1 & 0.9\cos^2(\frac{t\pi}{u}) & 0.6\sin^2(\frac{t\pi}{u}) \\ 0.5\sin^2(\frac{t\pi}{u}) & 0.4\sin^2(\frac{t\pi}{u}) & 0.1 & 0.9\cos^2(\frac{t\pi}{u}) \end{bmatrix}$$

For the first order Markov mixture model, we randomly choose 12 sets of initial values for the EM algorithm and iterate until a convergence criterion (Bohnig et al., 1994) is met or at most 500 iterations for each set of initial values. The one with the largest log-likelihood value is chosen as the final output.

In each simulation we generate N sequences from the transition matrices above. Each sequence has T observations. We compare the results of different sample sizes ($N = 50, 100$ for each cluster) and observation time points ($T = 50, 100, 200$). In the TM_{sc} model u is a tuning parameter that determines the speed with which TM_{sc} changes; the three cases we study are $u = 550$ when $T = 200$, $u = 300$ when $T = 100$, and $u = 150$ when $T = 50$. The accuracy of this procedure given in Table 5 is assessed for choosing the right number of clusters to see how well the Markov mixture model can cluster the observed sequences correctly. We first consider mixing three homogeneous clusters (TM_m, TM_s, TM_{sub}). Results of the model performance are shown in the upper panel in Table 1. Next, we try to mix homogeneous and non-homogeneous clusters for a three-cluster case and a five-cluster case. For each simulation we use 500 replications.

The simulation results show that when choosing the right number of clusters, with sufficient observations ($T = 100, 200$) this model is able to separate out the clusters with an accuracy above 90.0%, not only in the ideal situation ($> 96.0\%$) but also when mixing several non-homogeneous Markov clusters and a noisy cluster. Its performance appears to depend more on the number of observation time points rather than the number of sequences in a cluster. When there are clusters that do not fit the model assumptions, the accuracy of this clustering method seems to be more affected by an insufficient number of observations.

4.4 APPLICATION TO BIPOLAR YOUTH STUDY

The Course and Outcome of Bipolar Youth (COBY) study is a multicenter study being conducted at the University of Pittsburgh, Brown University and the University of California Los Angeles .Its aim is to characterize and prospectively follow youth with BD. In this study, 412 children and adolescents were followed between 6 months to 10 years and interviewed every 6 months were included. The average number of follow-up weeks is 340.8 ± 115.4. All youth were interviewed at intake using the Kiddie Schedule of Affective Disorders and Schizophrenia Present-Lifetime version (KSADS-PL) (Kaufman, et al. 2005) and over follow-up with the Longitudinal Interval Assessment Evaluation (LIFE) (Keller, et al. 1987). Week-by-week longitudinal change in psychiatric symptoms was assessed using the LIFE and quantified using the Psychiatric Status Rating (PSR) scale (Keller, et al. 1987). The PSR uses numeric values linked to the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV) criteria (American Psychiatric Association 2000). It has 6-point subscales for mania, hypomania, and depression. Based on clinical relevance, it is summarized into a 12-point mood rating score and a 4-point mood rating score (Table 6) for the analysis of the longitudinal pattern of youth BD. Examples of the observed sequences are given in Figure 4 and Figure 5. We use a heat map to show the observed sequences in Figure 6. Each row is an observed sequence, and different colors represent different mood states. (red – euthymic mood state; green – mania/hypomania mood state; yellow – depression mood state; blue – mixed mood state; white – missing). Our primary research interest is to learn about the mood transition patterns of children and adolescents with BD and cluster these subjects according to their patterns of mood changes. These findings could be instrumental for the understanding of the longitudinal course and treatment of BD and further research development.

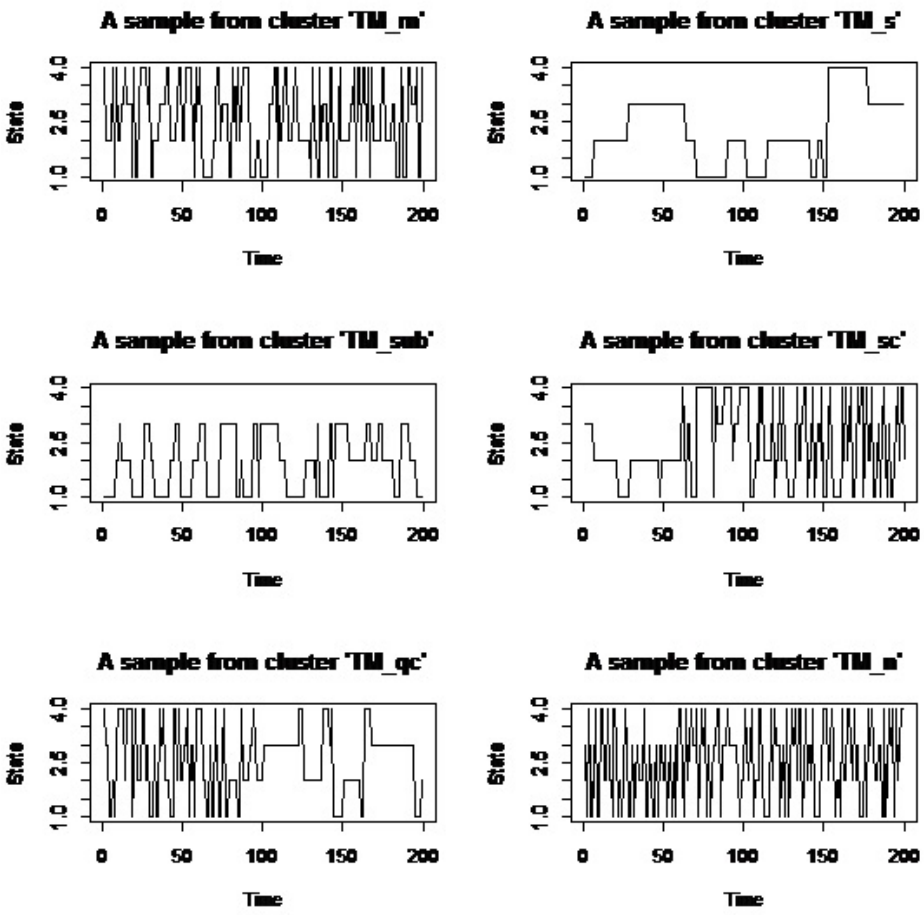

Figure 3: A simulation example in each cluster

Table 5: Simulation study for Markov mixture model when model assumptions fail

N	T	True clusters	average accuracy (sd)
100	200	TM_s, TM_m, TM_{sub}	100.0% (< .001)
100	100	TM_s, TM_m, TM_{sub}	99.8% (0.008)
100	50	TM_s, TM_m, TM_{sub}	98.1% (0.011)
50	200	TM_s, TM_m, TM_{sub}	99.9% (0.001)
50	100	TM_s, TM_m, TM_{sub}	99.6% (0.007)
50	50	TM_s, TM_m, TM_{sub}	96.6% (0.031)
100	200	TM_s, TM_m, TM_{sc}	99.9% (0.001)
100	100	TM_s, TM_m, TM_{sc}	99.5% (0.003)
100	50	TM_s, TM_m, TM_{sc}	96.5% (0.010)
100	200	TM_s, TM_m, TM_{qc}	99.7% (0.003)
100	100	TM_s, TM_m, TM_{qc}	96.9% (0.013)
100	50	TM_s, TM_m, TM_{qc}	87.1% (0.029)
100	200	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	97.3% (0.047)
100	100	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	93.7% (0.049)
100	50	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	80.2% (0.067)
50	200	TM_s, TM_m, TM_{sc}	99.9% (0.001)
50	100	TM_s, TM_m, TM_{sc}	99.4% (0.005)
50	50	TM_s, TM_m, TM_{sc}	95.8% (0.017)
50	200	TM_s, TM_m, TM_{qc}	99.6% (0.005)
50	100	TM_s, TM_m, TM_{qc}	96.9% (0.016)
50	50	TM_s, TM_m, TM_{qc}	86.0% (0.042)
50	200	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	96.2% (0.063)
50	100	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	92.1% (0.069)
50	50	TM_s, TM_m, TM_{sc}, TM_n , TM_{qc}	78.5% (0.065)

Previous work (Lopez, 2008) using the first two years of this dataset with a mixture Markov model to cluster the sequences found subgroups of mood-change patterns. Then HMMs were applied to incorporate covariates, such as age and gender, to identify factors that had effects on the mood transition. Recent work of Birmaher and his colleagues (2014) applied latent growth curve models to the COBY data and found four well-separated clusters. Accordingly, we believe that the mood change patterns among BD youth are heterogeneous. It is interesting and important to investigate these results from different perspectives in order to have a better understanding of how they differ.

In our study we analyze up to 10 years longitudinal data, and we are cautious at the clustering stage. Thus, before clustering we pay special attention to the model selection methods for choosing the optimal number of clusters and number of variables. Then we consider both semi-parametric (i.e. HMMs clustering algorithm) and parametric (i.e. MxM) methods for clustering the mood trajectories of the BD youth. Our goal is to find the well-separated and interpretable clusters, which would lay down a better foundation for including explanatory variables within clusters.

We apply the first order Markov mixture model to study the transition pattern among these four mood states and use AIC and BIC to choose the number of clusters. Both measures decrease as the number of clusters increases. The BIC shows the biggest jump between two and three clusters, then decreases more gradually for larger numbers of clusters, thus we chose the three-cluster solution. The largest cluster contains 248 subjects. The mood states of subjects in this cluster are fairly stable. The estimated probabilities of remaining in current mood state in the next observed time point are all above 0.9. The next largest cluster has 119 subjects. The probabilities of remaining in the current mood states reduce to 0.75 for those unhealthy mood states (i.e. depressed, maniac, mixed). The smallest cluster has 45 subjects. Mood fluctuations are much more common in this cluster than the first two. These three clusters are presented in Table 7 and Figure 7. Apparently, the smallest cluster on the right appears more serrated, while the largest cluster on the left shows more blocked coloring.

In fact, from the visualization of the largest cluster we notice there are in the main two different colors: red and blue. To further reduce heterogeneity, we carry out a hybrid

Table 6: Mood rating scale in bipolar youth study

Depression	Mania	Hypomania	12-point Score	4-point Score
$5-6$	1	$1-2$	1 : MDD* - Pure	3 : Pure Depression
$3-4$	1	$1-2$	2 : Subdepression only	3 : Pure Depression
$1-2$	1	$1-2$	3 : Euthymic (well)	1 : Euthymic (well)
$1-2$	1	$3-4$	4 : Submania only	2 : Pure mania/hypomania
$3-4$	1	$3-4$	5 : Submania / Subdepression	4 : Mixed
$5-6$	1	$3-4$	6 : Submania / MDD	4 : Mixed
$1-2$	1	$5-6$	7 : Hypomania - Pure	2 : Pure mania/hypomania
$3-4$	1	$5-6$	8 : Hypomania / Subdepression	4 : Mixed
$5-6$	1	$5-6$	9 : Hypomania /MDD	4 : Mixed
$1-2$	$5-6$	$1-2$	10 : Mania - Pure	2 : Pure mania/hypomania
$3-4$	$5-6$	$1-2$	11 : Mania / Subdepression	4 : Mixed
$5-6$	$5-6$	$1-2$	12 : Mixed state	4 : Mixed

* MDD is short for Major Depressive Disorder.

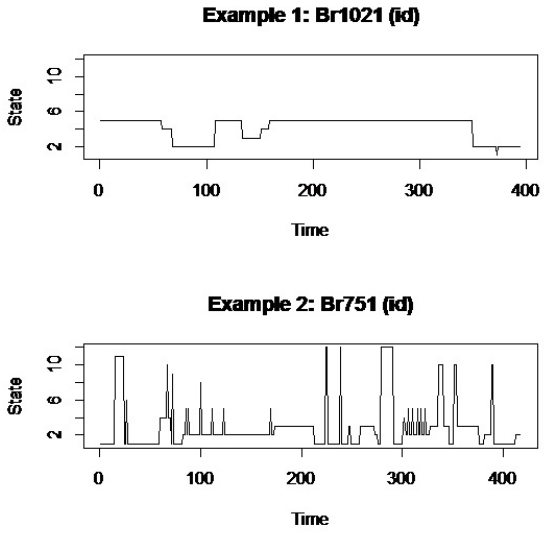

Figure 4: Example sequences (12-point rating score)

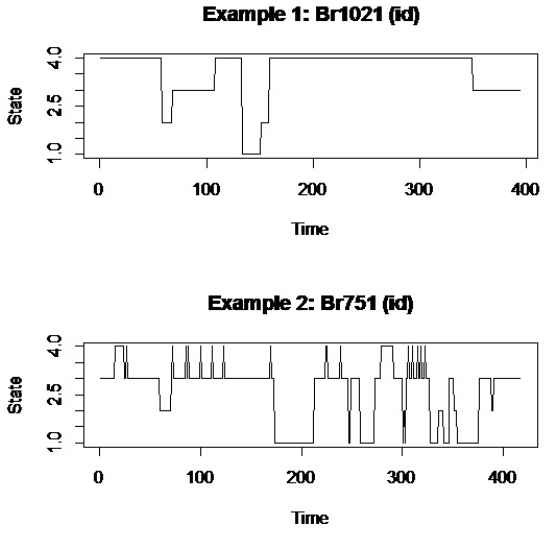

Figure 5: Example sequences (4-point rating score)

clustering scheme to further split the largest cluster according to the percentage of time a subject stays on certain mood state, using the K-means algorithm. As a result, we find two subgroups within the largest cluster. One has 125 subjects who are most time in euthymic mood. The average percentage of time in euthymic mood state is 75.9%. The other has 123 subjects who are "mostly mood symptomatic": 38.0% of time in mixed mood state, 27.3% of time in depressed mood and 16.1% of time in manic/hypomanic mood state, on average. The two subgroups are displayed in Figure 8.

We conclude that BD children and adolescents are not a homogeneous group. More than half of them remain in a relatively stable mood for a long time. Only around 10% of BD youth have mood switches at a high frequency. Since the transition patterns appear to be different among clusters, follow-up studies, such as including covariates to explain differences in transition matrices, should be conducted within clusters.

4.5 DISCUSSION

To better understand the recurrent mood fluctuations usually observed in bipolar disorder, in this chapter we introduced the Markov mixture model to cluster discrete sequences. Unlike traditional longitudinal models focusing on trends over time, this method provides a novel angle to analyze longitudinal data in a mental health study, which is to characterize the discrete sequences based on their state transition patterns. The model demonstrates a satisfactory and stable performance in the simulation study, even when some of the model assumptions do not hold. In the BD application, we find that the mood transition patterns among children and adolescents with BD are heterogeneous; in fact, there are three well separated more homogeneous clusters based on mood transition patterns. To enhance homogeneity, we adopt a hybrid clustering scheme, using a K-means algorithm to further split the largest cluster according to the percentage of time in each mood state. As a result, we separate out those children and adolescents who are in euthymic mood state most of the time from those are more often mood symptomatic, within the largest stayer cluster.

However, there are several challenges in the application of first-order Markov mixture models to BD clustering problem. First, the cluster assignments may change when we start with different initial values in the EM algorithm. When the number of clusters increases to four or more, the algorithm gets trapped in local maxima, which is common in mixture model clustering algorithms. Thus, we regard this instability as a useful complement to the AIC or BIC criteria for choosing the number of clusters. Second, some observed sequences show departures from time homogeneity of a Markov chain by a chi-square test (Anderson, et al. 1957; Bianca, et al. 1988). In the simulation, we observe that a Markov mixture model can separate the time-inhomogeneous Markov chains, even though it does not detect the change in the transition matrices. We intend to study this phenomenon next by including explanatory variables which may help relate the mood state changes over time and by those factors. Third, the estimated transition matrices may not always represent the transition patterns well. For instance, when the cluster consists of time-inhomogeneous Markov chains, the transition matrix would change over time and cannot be reflected in one estimated transition matrix. Consider the smallest cluster (the most frequently mood-change cluster)

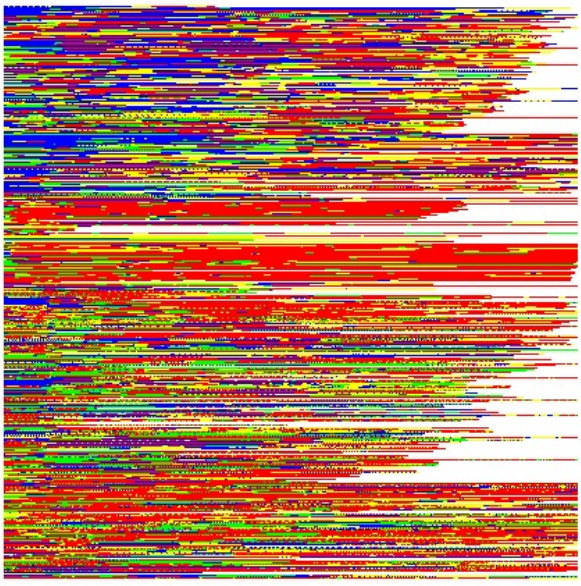

Figure 6: Weekly mood rating of 412 children and adolescents with bipolar disorder.

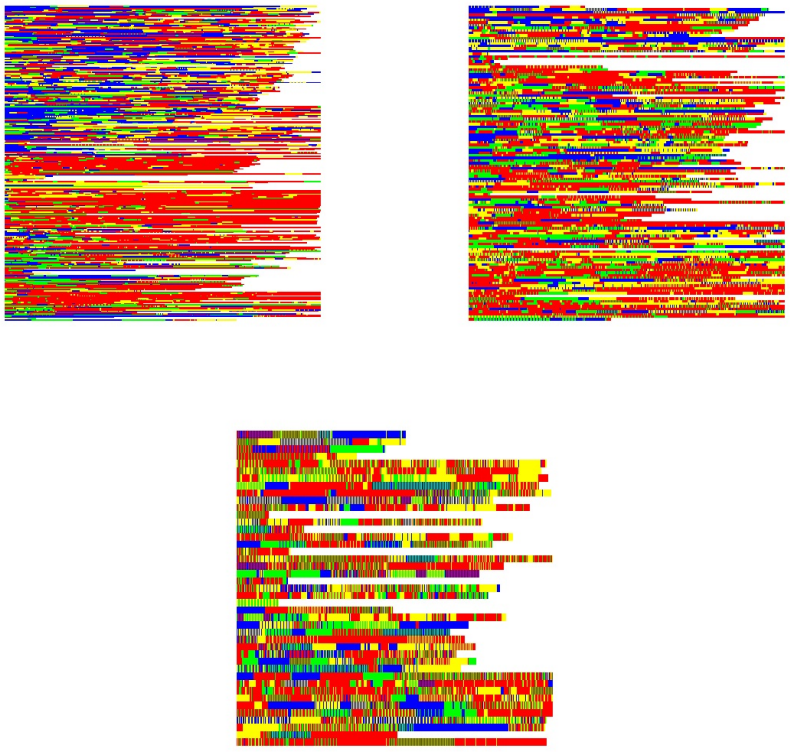

Figure 7: Markov mixture model separates three clusters in bipolar youth study

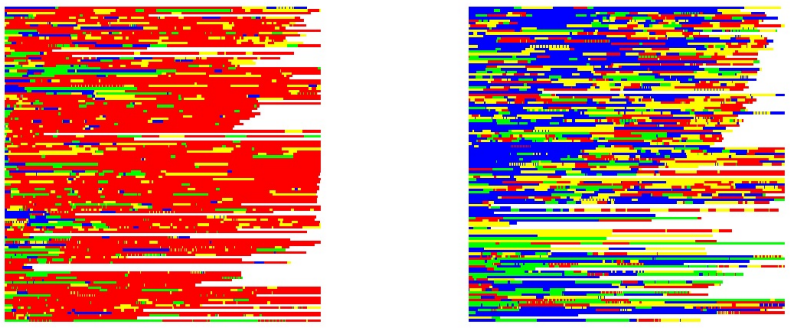

Figure 8: Splitting the largest cluster in bipolar youth study

Table 7: The estimated transition matrix in bipolar youth study

Cluster 1 ($n = 248$)	Cluster 2 ($n = 119$)	Cluster 3 ($n = 45$)
$\begin{bmatrix} 0.97 & 0.01 & 0.02 & 0.00 \\ 0.04 & 0.93 & 0.01 & 0.02 \\ 0.05 & 0.00 & 0.93 & 0.02 \\ 0.01 & 0.01 & 0.02 & 0.96 \end{bmatrix}$	$\begin{bmatrix} 0.87 & 0.07 & 0.05 & 0.01 \\ 0.21 & 0.71 & 0.03 & 0.05 \\ 0.10 & 0.03 & 0.76 & 0.11 \\ 0.03 & 0.05 & 0.16 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 0.68 & 0.17 & 0.10 & 0.05 \\ 0.36 & 0.35 & 0.15 & 0.14 \\ 0.20 & 0.15 & 0.53 & 0.13 \\ 0.13 & 0.18 & 0.18 & 0.51 \end{bmatrix}$

in the BD study for another example. We find that even though all the sequences in this cluster switch mood frequently, the change "preferences" vary: some change from mixed to euthymic mood states more often while others change from depressed to mixed mood states more often. Hence, we should be more cautious to label the clusters simply based on the estimated model parameters. Visualization tools, such as a heat map, are useful to get a better sense of the characteristics of clusters beyond estimated transition matrices. Further inference and investigation are needed for more compelling conclusion. Last but not least, existing methods for testing Markov assumption [19] are based on a stationarity assumption, which need not hold in the BD study. In this paper, we focus on the first-order switch patterns, since it has a clear clinical sense. However, we do not rule out the possibility that there may be higher-order connections. In sum, clustering usually is an exploratory stage in a study. In the next step, an interesting question for psychiatrists and researchers is what explains the mood change patterns in each cluster.

BIBLIOGRAPHY

[1] Aguirre-Hernandez, Farewell VT. (2002) A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine*, 21, 1899–1911.

[2] Ait-Sahalia Y., Fan J., Jiang J. (2010) Nonparametric tests of the Markov hypothesis in continuous-time models. *The Annals of Statistics*, 38: 3129-3163.

[3] Altman RM. (2004) Assessing the goodness-of-fit of hidden Markov models. *Biometrics*, 60, 444–450.

[4] Altman RM. (2007) Mixed midden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102.

[5] Anderson T., Goodman L. (1957) Statistical inference about Markov chain. *The Annals of Mathematical Statistics*, 28, 89-110.

[6] Azzalini A. (1996) Statistical Inference based on the likelihood. Chapman and Hall.

[7] Bianca DS. (1988) Testing departures from time homogeneity in multistate Markov processes. *Journal of the Royal Statistical Society*, 37, 242-250.

[8] Bicego M., Murino V., Figueiredo A. (2003) Similarity-based clustering of sequences using hidden Markov models. *Machine Learning and Data Mining in Pattern Recognition*, 2734, 86-95.

[9] Bickel P, Ritov Y. (1996) Inference in hidden Markov models: local asymptotic normality in the stationary case. *Bernoulli*, 2, 199–228.

[10] Bickel P, Ritov Y, Ryden T. (1998) Asymptotic normality of the maximum-likelihood estimator for general Hidden markov models. *Annals of Statistics*, 26, 1614–1635.

[11] Bilmes JA. (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4 (510), 126.

[12] Bohnig D, Dietz E, Schaub R, Schlattman P, Lindsay BG. (1994) The distribution of the likelihood ratio for mixtures of densities from the one parameter exponential family. *Annals of the Institute of Statistical Mathematics* , 46, 373-388.

[13] Cadez I, Heckerman D, Meek C, White S. (2003) Model-based clustering and visualization of navigation patterns on web site. *Data Mining and Knowledge Discovery*, 7, 399-424.

[14] Camproux A., Saunier F., Chouvet G., Thalabard J., Thomas G. (1996) A hidden Markov model approach to neuron firing patterns. *Biophysical Journal*, 71, 2404-2412.

[15] Cavanaugh J, Neath A. (1999) Generalizing the derivation of Schwarz information criterion. *Communications in Statistics - Theory and Methods*, 28, 1, 49-66.

[16] Chen B, Hong Y. (2012) Testing for the Markov property in time series. *Econometric Theory*, 28, 130-178.

[17] Crayen C, Eid M, Lischetzke T, Courvoisier D, Vermunt J. (2012) Exploring Dynamics in Mood Regulation: Mixture Latent Markov Modeling of Ambulatory Assessment Data. *Psychosomatic Medicine*, 74, 366-376.

[18] Coetzee FM. (2005) Correcting the Kullback-Leibler distance for feature selection. *Pattern Recognition Letters*, 26, 1675–1683.

[19] Coviello E, Chan A, Lanckriet G. (2012) Clustering hidden Markov models with variational HEM. *Journal of Machine Learning Research*, 1 - 44.

[20] Fahrmeir L, Kaufmann H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13, 342-368.

[20] Feng D, Tierney L, Magnotta V. (2012) MRI tissue classification using high-resolution Bayesian hidden Markov normal mixture models. *JASA*, 107, 102–119 .

[21] Fung P, Ngai G, Cheung C-S. (2003) Combining optimal clustering and hidden Markov models for extractive summarization, *Proceedings of the ACL*, 21–28.

[22] Gales M, Young S. (1992) An improved approach to the hidden Markov model decomposition of speech and noise. *Acoustics, Speech, and Signal Processing*, 1, 233-236.

[23] Garcia-Garcia D, Parrado-Hernandez E, Diaz-de-Maria F. (2011) State-space dynamics distance for clustering sequential data. *Pattern Recognition*, 44, 1014-1022.

[24] Hershey J, Olsen P, Rennie S. (2007) Variation Kullback-Leibler divergence for hidden Markov models. *IEEE Workshop on Automatic Speech Recognition & Understanding.*

[25] Hilton G. (1971) An algorithm for detecting difference between transition probability matrices. *Journal of the Royal Statistical Society*, series C, 20, 80-86.

[26] James G, Sugar C. (2003) Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, 98, 397-408.

[27] Jebara T, Song Y, Thadani K. (2007) Spectral clustering and embedding with Hidden Markov models. *Machine Learning: ECML*, 4701, 164–175.

[28] Jebara T, Kondor R. (2003) Bhattacharyya and expected likelihood kernels. *Learning Theory and Kernel Machines*, Springer.

[29] Johansson M, Olofsson T. (2007) Bayesian model selection for Markov, hidden Markov, and multinomial models. *IEEE Signal Processing Letters*, 14, 129–132.

[30] Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D, Ryan N. (1997) Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36, 980-988.

[31] Keller M, Lavori P, Friedman B, Nielsen E, Endicott J, McDonald-Scott P, Andreasen N. (1987) The longitudinal interval follow-up evaluation: a comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry*, 44, 540.

[32] Komarek A, Komarkova L. (2013) Clustering for Multivariate Continuous and Discrete Longitudinal Data. *The Annals of Applied Statistics*, 7, 177-200.

[33] Krogh A, Larsson B, Heijne G. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305, 3, 567-580.

[34] Louis T. (1982) Finding the observed information matrix when using EM algorithm. *Journal of the Royal Statistical Society* Series B, 44, 226–233.

[35] Lopez A. (2008) Markov models for longitudinal course of youth bipolar disorder. PhD thesis, Statistics Department, University of Pittsburgh.

[36] Luxburg U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, Springer.

[37] Mamon, ER. (2010) Hidden Markov Models in Finance. *International Series in Operations Research and Management Science*, Springer.

[38] McCulloch CE, Searle SR, Neuhaus JM. (2008) *Generalized, Linear, and Mixed Models* (2nd edition) , Wiley.

[39] Meredith W, Tisak J. (1990) Latent curve analysis. *Psychometrika* , 55, 107122.

[40] Molenberghs G, Verbeke G. (2008) *Models for Discrete Longitudinal Data*. Springer.

[41] Ng A, Jordan M, Weiss Y. (2002) On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14.

[42] Nwulia EA, Zandi PP, McInnis MG, DePaulo JR Jr., MacKinnon DF. (2008) Rapid switching of mood in families with familial bipolar disorder. *Bipolar Disorder*, 10, 597-606.

[43] Panuccio A, Bicego M, Murino V. (2002) A Hidden Markov Model-based approach to sequential data clustering. *SSPR/SPR*, 734–743.

[44] Rao V, Rao C, Yeragani V. (2006) A novel technique to evaluate fluctuations of mood: implications for evaluating course and treatment effects in bipolar/affective disorders. *Bipolar Disorder*, 8, 453-466.

[45] Rabiner L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, 77, 2.

[46] Ait-Sahalia Y, Fan J, Jiang J. (2010) Nonparametric tests of the Markov hypothesis in continuous-time models. *The Annals of Statistics*, 38, 3129-3163.

[47] Scott SL, James GM, Sugar CA. (2005) Hidden Markov models for longitudinal comparisons. *JASA*, 100, 359–369.

[48] Shorrock A. (1978) The measurement of mobility. *Econometrica*, 46, 1013-1024.

[49] Smith A, Naik PA, Tsai C-L. (2006) Markov-switching model selection using Kullback-Leibler divergence. *J Econometrics*, 134, 553–577.

[50] Smyth P. (1997) Clustering sequences with hidden Markov models. *Advances in Neural Information Processing Systems*. 9.

[51] Soding J. (2004) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21, 951-960.

[55] Sung M, Soyer R, Nhan N. (2007) Bayesian analysis of non-homogeneous Markov Chains: application to mental health data. *Statistics in Medicine*, 26, 3000-3017.

[56] Thacker N, Aherne F, Rockett P. (1997) The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34, 4, 363-368.

[57] Titman AC, Sharples LD. (2008) A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, 27, 2177–2195.

[58] Turner TR, Cameron MA, Thomson PJ. (1998) Hidden Markov chains in generalized linear models. *Canadian Journal of Statistics*, 26, 107–125.

[59] Vidyasagar M. (2007) Bounds on the Kullback-Leibler divergence rate between hidden Markov models. *IEEE Conference on Decision and Control*, 6160–6165.

[60] Visser I, Raijmakers MEJ, Molenaar PCM. (2002) Fitting hidden Markov models to psychological data. *Scientific Programming*, 10, 185–199.

[61] Zhu H, He Z, Leung H. (2012) Simultaneous feature and model selection for continuous hidden Markov models. *IEEE Signal Processing Letters*, 19, 279–282.