THE UNIVERSITY OF CHICAGO

CLASSIFICATION OF SLEEP STAGE BASED ON EEG WAVE

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF STATISTICS

BY

HONG XU

CHICAGO, ILLINOIS

JULY 2005

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank my master's thesis adviser, Professor Zhiyi Chi, for his guidance and support. I also want to thank all the faculty and students in the Department of Statistics at the University of Chicago for providing this great study opportunity and environment. Thanks are also due to Professor Daniel Margoliash in the Department of Organismal Biology and Anatomy and Department of Psychology for sharing the data and his graduate student Sylvan Shank for data collection and providing background information on his experiment.

My deepest gratitude goes to my husband, Wanli Min, and my family. It is hard to imagine that this work could ever have been done without their support and encouragement.

Abstract

Electroencephalogram (EEG) has been used to show the electrical activities of the brain, and therefore it has been proved a powerful channel to obtain latent brain state of subjects. In this work, we try to extract the information on the sleep stage of birds by the measured EEG wave. We analyze the EEG signal in frequency domain by various techniques. In particular, we compare hidden Markov model (HMM) and a Mixed Gaussian model for the estimated spectral density in multiple frequency bands, by using maximum-a-posterior (MAP) criterion to classify the hidden state. The effectiveness of the approaches is compared in both simulations and application to the real data. Some related issues in model assessment are also discussed.

1 Introduction

Electroencephalogram (EEG) study dates back to Berger (1929), Empson (1989). Berger studied the electrical activity of the brain and the state of the brain by externally attaching electrodes on the human skull. This technique later was used in sleep study to investigate the state of sleep. Rapid eye movement (REM) sleep in human was first discovered by Aserinsky E. and Kleitman N. in 1953 at the University of Chicago, Aserinsky and Kleitman (1953). Similar to human sleep, bird sleep is also an interesting topic. Since early nineteenth century, studies of sleep patterns has been reported in pigeon. From the first field experiments on sleep at the mid of 1970s, sleep studies on birds expanded rapidly.

Basically there are two types of sleep for birds: quite sleep and active sleep. Quiet sleep is also called non-rapid eye movement sleep (NREM). NREM sleep has a synchronized, high-amplitude EEG wave (175 to 300μV) with a strong low-frequency component (2 to 5.5 Hz, Amlander and Ball (1994)). The median length of a NREM episode lasts 144 seconds.

Active sleep is also called rapid eye movement sleep (REM) and it accounts 5 to 10 per cent of total sleep time. Each episode of REM lasts about 9 seconds on average. Its EEG wave is characteristically desynchronized with relatively high

frequency ($10 - 23$ Hz) and low amplitude ($< 50\mu V$), Amlander and Ball (1994). The eyes are usually closed in this sleep stage, and arousal thresholds are high. The average REM time in one night is reported at 0.2 hour, Schmidt *et al.* (1990).

In general, NREM sleep has low frequency but high amplitude EEG wave, lasts longer than REM sleep, Amlander and Ball (1994). According to Tobler (2005), NREM sleep power density values in the low frequency range (0.25 - 6.0 Hz) exceed those of REM sleep by approximately one order of magnitude, which is also indicated in Figure 1.

Between these two sleep states, the bird may be awake or move. And some of these states may share the same EEG frequency feature as REM or NREM. One of such states is drowsiness, which shares similar features as Quite Wake and NREM and usually occurs in between these two states. Therefore, drowsiness is not always recognized as a separate state. One other state is called Cataleptic immobility and also called sleeplike state 1, in which bird has open eyes and reduces responsiveness. The amplitude of its EEG wave is high (200-300 μV) and has bimodal frequency (1-6 Hz and 8-12 Hz), Amlander and Ball (1994). The high amplitude and low frequency EEG with open eyes makes this state similar to NREM, or drowsiness or quite wake sometimes. And this state occupy 20 per cent of the total sleep time. Similarly,

Vigilant Sleep is a state hard to separate from NREM. It is an open-eye period that have desynchronized, relatively low-amplitude EEG signals, and its arousal threshold is between those of NREM and Quite Wake. Although this state is frequently scored as NREM, it should be treated cautiously for its arousal threshold and EEG.

The state similar to REM is called gaze wakefulness in which the bird is immobilized and opens or partly opens its eyes. Eyes move slowly and phasically, especially at the onset of each episode. Its EEG signal is desynchronized. But the difference between these two states is that eyes are closed and moving rapidly in REM, Amlander and Ball (1994).

To make a clear distinction in these intermediate states, the current analysis only includes the data with either NREM or REM states.

Compared to human or mammal sleep, birds has more sleep states. This may be related to the shorter episode length of sleep stages, which in turn forces researchers to look more closely at shorter time intervals. By looking at the trend of REM and NREM in sleep time, researchers found that birds have similar amount of NREM sleep time of the mammals per day but only have about 1/4 of the amount of REM of the mammals. It is reported that perching birds have extremely small

amount of REM (0.05 to 0.17 h per 24 hours), Amlander and Ball (1994). Schmidt *et al.* (1990) found the average REM time was 0.2 hour per 24 hours. Also the distribution of REM and NREM are slowly changing over night. NREM starts rapidly at the onset of sleep, and it remains dominant throughout the sleep period. Whereas REM shows up slowly, but rises rapidly to peak at about three quarters of the sleep period and then declines quickly around dawn, Amlander and Ball (1994).

Traditionally the sleep stage is manually classified based on the EEG waves together with eye movements and muscle tension recorded by Electrooculogram (EOG) and Electromyogram (EMG) respectively. In the current data set, the researchers used EEG together with the video which recorded the behaviors of the birds during sleep. However, it takes a long time to manually go through the whole sleep sequence and classify the states because of the large quantity of EEG data, the randomness of the EEG signals and the short-period states of birds. Besides, subjective criteria may vary between different scorers. Therefore an objective classification of sleep states is strongly needed. The previous study by Nick and Konishi (2001) used some Matlab functions to classify the slow wave sleep (NREM), but it focused on slow wave sleep classification. They also pointed out the urgency of using objective criteria for sleep stages. In this work, the purpose is to develop an objective method to classify the sleep states: NREM and REM in particular.

There are enormous literature on modelling EEG signals and classifying the underlying states. A short list of them is given here. More details can be found through reference therein. Sergejew and Tsoi (1996) used Markov modelling of an AR representation of the EEG signal to quantify EEG state transition. They found limited state transition dynamics in the EEG of Obsessive-Compulsive Disorder (OCD) patients but not in that of normal subjects. Cohen *et al.* (1996) segmented the non-stationary vector EEG signal into stationary records, by using a vector AR(6) segmentation algorithm. Then they classify each segment into a sleep state, using a nearest neighbor classifier with Kullback-leibler based distortion measure (Gersch *et al.* (1979)). The average correctness over four human patients is about 85%. Gersch (1996) proposed the time varying autoregressive (TV-AR) coefficient model to model the time series of scalar nonstationary covariance EEG wave. Using stochastic partial correlation coefficient (PARCOR) model on segments of seizure episodes in human, Gersch found that the abrupt change in the power spectral density is better captured by Cauchy noise than Gaussian noise in the AR model (also see Kitagawa (1987)). Mutapcic *et al.* (2003) used two feature extraction methods: classical fast Fourier transform (FFT) analysis and least-mean squares (LMS) based feature extraction, and then used a 2-layer neural network for classifying sleep stages of patients' EEG data. They found similar overall (high 70% to low 80%) and

per-stage (mid 70% range) classification accuracy in both methods. The correctness of classification on the same patient as training is about 80%, and about 70% on the patients different from training.

Hidden Markov model (HMM) is used by Penny and Roberts (1998) on simulated Gaussian observation generated on AR or MAR coefficients. By applying directly to the frequency data, they trained HMM by EM algorithm and classify the states by Viterbi decoding. They found "HMM can detect changes in DC levels, correlation, frequency and coherence that are typical of the nonstationarities in an EEG signal". Also they pointed out that cluster analysis of derived features in the data can be used to choose the number of hidden states in HMM.

To estimate parameters of hidden Markov model is a challenging problem. Maximum likelihood or maximum *a posterior* is frequently adopted. The optimization is often carried by EM algorithm. Andrieu and Doucet (2000) used simulated annealing with data augmentation. They also proved that under certain constraints, the Markov chain generated by the simulated annealing (SA) converges. They applied this method to data simulated by an AR process with a Markov regime. By comparing the results of SA algorithm with EM algorithm, they found the results are similarly good: the estimates' bias and variance with respect to the

true value are very small. But SA performs better than EM when they are applied to data generated by Markov modulated Poisson processes. Even though SA can find global optimizer under certain conditions, its considerably slow speed and too many tuning parameters in implementation makes it difficult to use. On the other hand, EM algorithm converges significantly faster than SA. To avoid possible local minimum, EM algorithm can be executed with different initial parameter values.

The rest of this paper is organized as follows. Section 2 describes preprocessing of the EEG data. Section 3 proposes the nonhomogeneous hidden Markov model followed by a simulation study in Section 4. Section 5 presents the results of the models on real data and discusses related issue. Section 6 concludes.

2 Data collection and feature extraction

The raw data was collected on individual bird (zebra finch), each with one or two whole nights' sleep (8-hour period). Single channel EEG waves were recorded at the forbrain of the bird, either left or right hemisphere of the brain, by attaching two electrodes between the skull and dura. One electrode is served as the reference.

The raw data is time series of EEG recordings at sampling rate 1KHz. Because of the large amount of data, feature extraction is needed to capture the useful

information embedded in the raw data. To this end, we consider the power spectral density. In fact, it has been argued by many researchers that EEG signal collected within $1\sim$ 2 seconds can be represented by autoregressive model (AR) of certain order, typically $5 \sim 7$, Haselsteiner and Pfurtscheller (2000), Anderson *et al.* (1995). Penny and Roberts (1999) apply a Kalman filter (KF) to obtain the AR coefficients at each time point. On the other hand, most time series processes have their equivalent representations in frequency domain, including AR(p) processes, see Brockwell and Davis (1991). Second, the PSD of REM/NREM of our data is clearly distinguishable. Fig. 1 shows that the average PSD within lower frequency range (below 30HZ) is lower in REM stage than that of NREM stage. We believe the PSD within lower frequency range can discriminate REM/NREM states and will rely on this criteria to label the state space of the hidden Markov chain to be discussed soon.

For each recording, its power spectral density is obtained using the Thompson multi-taper method, Thompson (1982). Time shift of moving windows in this method is 0.3s with duration of 3s each. So in total, we have 96000 data points for one night's sleep time. Because the power of EEG is concentrated in frequencies below 40 HZ ("low passed", see Fig.1), only power spectral density in frequencies up to 60HZ is to be used. For each frequency, normalize the logarithm of PSD across time to get Z scores (subtract mean and divided by standard deviations). Let this

matrix be M_{ft}. Based on this matrix, we derive multi-band scores as follows.

First, choose several non-overlapping frequency bands in the range of 1HZ $\sim$ 60HZ. For example, bands 1HZ $\sim$ 5HZ, 5.5HZ $\sim$ 10HZ, 10.5HZ $\sim$ 20HZ, 20.5HZ $\sim$ 30HZ. In our project, we will only consider $3 \sim 4$ bands. Then we pick the largest Z score among all these frequencies in cth band at each time point. That is, for the cth band, at each time point t, let

$$\widetilde{X}_{ct} = \max\{Z_{ft}, \text{ for all frequnices } f \text{ in band } c\}. \tag{1}$$

Because occasionally the recording has very large noise caused by movement, we truncate the above score by

$$\text{sign}(X_{ct}) \min\{A, |X_{ct}|\}, \tag{2}$$

still denoted by X_{ct}. Here we set A = 5. For example, if $X_{ct} = 2$, then it is not changed after the truncation. If $X_{ct} = -A - 1$, it becomes $-A$ after the truncation. The random vector at time t then consists of $\boldsymbol{X} = (X_{1t}, \ldots, X_{Kt})$. K is the total number of bands. The classification will be based on the observations $\boldsymbol{X}_t$ across time.

Fig. 2 plots the probability density of multi-band scores estimated from one epoch of the time series $\{\boldsymbol{X}_t\}$. A multi-modal characteristics is clear. Therefore it is natural to model $\boldsymbol{X}_t$ by mixture models. On the other hand, Fig. 3 and Fig. 4 shows strong

serial correlations and non-stationarity. So a hidden Markov model (HMM), Rabiner (1989), whose marginal distribution of observations exhibits mixture behavior and observations are correlated through a latent Markovian process, becomes another natural approach.

Despite their popularity in EEG analysis, mixture models and HMM are to some extent oversimplified. In fact, mixture model ignores the dependence over time indicated by Fig. 3 and Fig. 4. Traditional HMM assumes the transition probability between latent states as constant (homogeneous), i.e. $P(S_{t+1} = s|S_t = v, \ldots, S_1, \boldsymbol{X}_t, \ldots, \boldsymbol{X}_1) = P(S_{t+1} = s|S_t = v) = q_{vs}$, which is fundamentally flawed. It is like assuming an exponential distribution of lifetime in reliability studies, which often is to the contrary of the truth. In sleep stage study, it is not the case the probability of evolving from REM to NREM stays the same regardless of how long the bird has been in REM state. The ratio of REM to NREM starts very low, and increases slowly over night and reaches the peak rapidly at around three quarters of the sleep period, Amlander and Ball (1994). Taking these into account, we proposed a first order nonhomogeneous hidden Markov model (NHMM) whose state transition probability matrix is time dependent and modulated by some covariates $\boldsymbol{Z}_t$, i.e. $P(S_{t+1} = s|S_t = v, \ldots, S_1, \boldsymbol{X}_t, \ldots, \boldsymbol{X}_1, \boldsymbol{Z}_{t+1}, \ldots, \boldsymbol{Z}_1) = P(S_{t+1} = s|S_t = v, \boldsymbol{X}_t, \boldsymbol{Z}_{t+1}) = q_{vs}(t)$. The covariates $\boldsymbol{Z}_t$ could be additional variables which carry information of the current

state, for instance EEG signal through other channels from different locations, or they may just be the past observations $\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_1$. Therefore, our NHMM model has two merits. First, it can easily incorporate information conveyed by other variables through $\boldsymbol{Z}_t$, for instance, through polytomous logistic link function (See McCullagh and Nelder (1989)). Second, many physical processes assert that transitions from one state to another depends on the current observable. For example, one fundamental theory of LASER asserts that the quantum system can jump from state of energy level E_i to E_j with $P(i \to j) \propto e^{-(E_j - E_i)/KT}$. In view of the fact that EEG signal is related to some real physical processes of birds, it is natural to believe a model which models the process in first principle makes more sense, and that model is our NHMM.

3 General models

3.1 Multivariate nonhomogeneous hidden Markov model

We state our self-exciting nonhomogeneous hidden Markov model (NHMM) with conditional multivariate Gaussian observations as follows:

$$\begin{aligned} P(S_{t+1} = s | S_t = v, \cdots, S_1, \boldsymbol{X}_t, \cdots, \boldsymbol{X}_1) &= P(S_{t+1} = s | S_t = v, \mathbf{X}_t) := q_{vs}(t) \\ &= e^{\lambda_{sv} + \mathbf{q}_s^t \mathbf{X}_t} / \sum_{\xi=1}^{H} e^{\lambda_{\xi v} + \mathbf{q}_\xi^t \mathbf{X}_t} \qquad (3) \\ P(\mathbf{X}_t | S_t = s, \cdots, S_1, \boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_1) &= P(\boldsymbol{X}_t | S_t = s) \sim N(\mu_s, \boldsymbol{\Sigma}_s) \qquad (4) \end{aligned}$$

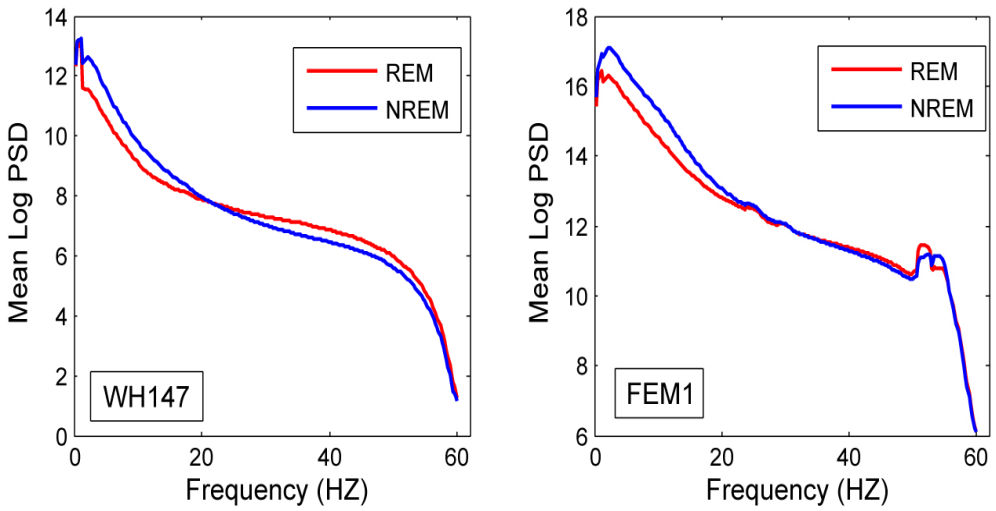

Figure 1: Log power spectrum density of EEG signal from two birds (WH147 and FEM1). The PSD is estimated for each epoch of length 3 seconds and averaged over the same sleep stages.

$$\lambda_{1v} = 0,\ \boldsymbol{q}_1 = \mathbf{0},\ P(S_1 = s) = \pi_s,\ v, s \in \{1, \ldots, H\},\ t = 1, \ldots, T \tag{5}$$

The parameters λ_{1v}, $\boldsymbol{q}_1$ are set to zero to guarantee identifiability of the transition probability parameters. Homogeneous HMM is a special case of NHMM with $\boldsymbol{q}_s = \mathbf{0}$ for all $s \in \{1, \ldots, H\}$. Fig. 5 also illustrates the difference between NHMM and HMM. The analysis of NHMM consists of two stages, parameter estimation and classification of the unobserved state sequence $\{S_t\}, t = 1, \cdots, T$, and both can be derived in a similar way as HMM for the key assumption of Markovian properties. We present major formulas here and leave details to the interested readers. In the sequel Θ stands for the parameters $(\pi, \lambda, \boldsymbol{q}, \mu, \Sigma)$ of the NHMM and we suppress subscripts whenever

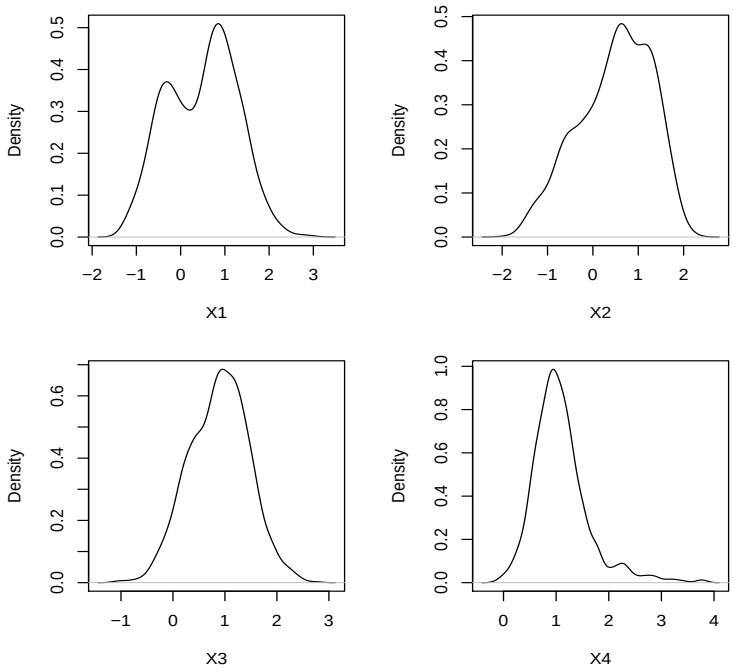

Figure 2: Estimated marginal probability density of the multi-band scores of log PSD of training sample D from bird WH147. X_1 through X_4 correspond to frequency band 1HZ $\sim$ 5HZ, 5.5ZH $\sim$ 10HZ, 10.5HZ $\sim$ 20HZ, 20.5HZ $\sim$ 35HZ.

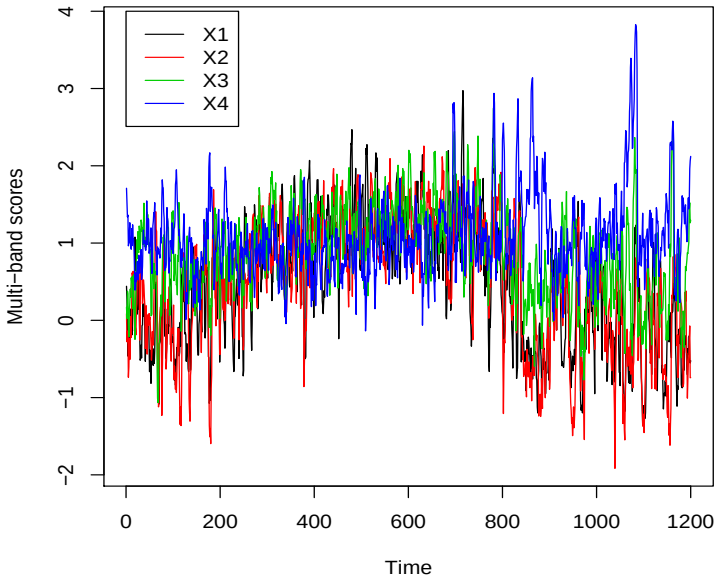

Figure 3: Time series plot of $\{\boldsymbol{X}_t, t = 1, \ldots, 1200\}$, the multi-band scores of log PSD of training sample D from bird WH147. X1 to X4 are four bands as in Figure 2.

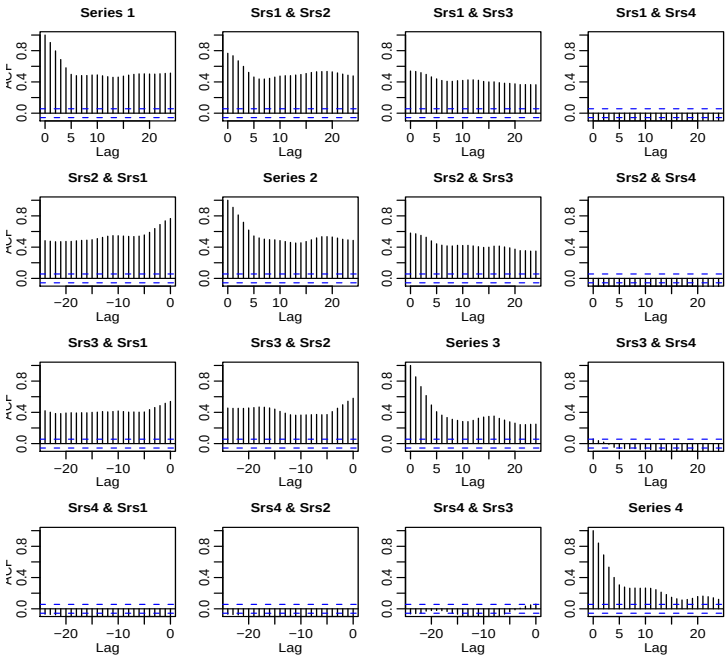

Figure 4: Autocorrelation function of $\{\boldsymbol{X}_t, t = 1, \ldots, 1200\}$, the multi-band scores of log PSD of training sample D from bird WH147. Series 1 to 4 are four bands as in Figure 2.

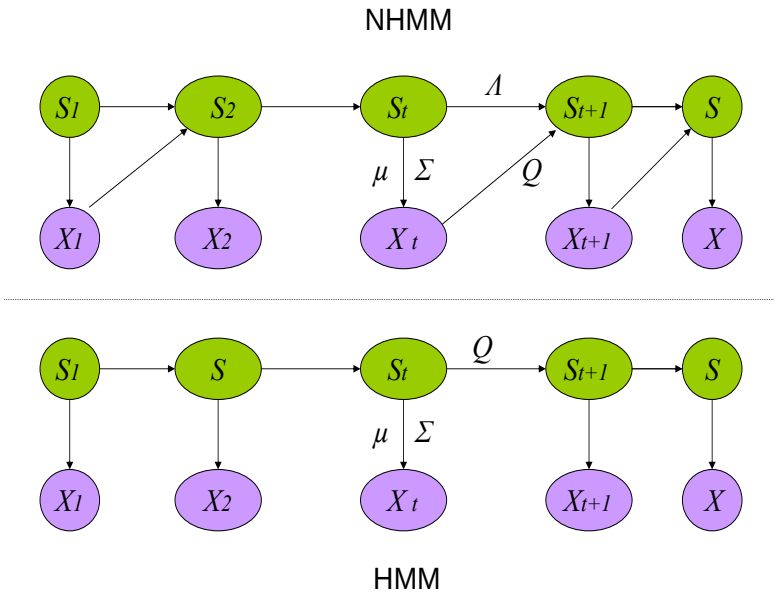

Figure 5: Graphical representation of NHMM and HMM, where S_t are the hidden state sequence.

possible.

3.2 Train NHMM by EM algorithm

EM algorithm is well suited to obtain the maximum likelihood estimate (MLE) of the parameters in the case of missing observations or existence of latent variables, Dempster *et al.* (1977). Despite its susceptibility to local maxima, EM algorithm has been applied successfully in many research areas, including estimation of model parameters of HMM. The pioneer Baum-Welch algorithm, Baum *et al.* (1970), also known as Forward-Backward algorithm, originally proposed to train HMM with discrete observations, is actually a variant of EM algorithm. It is readily to generalize to the case of HMM with continuous observations. In the sequel, I will give a brief account of this algorithm considering two cases, training sample consisting of either one sequence or several *independent* sequences governed by the *same* NHMM model parameters.

Let's start with training sample of single sequence. We define several auxiliary variables, for any $s \in \{1, \cdots, H\}$, Forward variable (FV) $\alpha_t(s) = P(S_t = s, \mathbf{X}_1, \cdots, \mathbf{X}_t|\Theta)$, Backward variable (BV) $\beta_t(s) = P(\mathbf{X}_{t+1}, \cdots, \mathbf{X}_T|S_t = s, \boldsymbol{X}_t, \Theta)$ for $t = 1, \cdots, T$ with the understanding of $\beta_T(s) = 1$. By Markovian property, the following recursion equations hold:

$$\alpha_t(s) = \begin{cases} \pi_s f(\mathbf{X}_1|S_1 = s, \Theta), & t = 1 \\ \sum_{v=1}^{H} \alpha_{t-1}(v) q_{vs}(t-1) f(\mathbf{X}_t|S_t = s, \Theta), & t = 2, \cdots, T \end{cases} \tag{6}$$

$$\beta_t(s) = \begin{cases} 1, & t = T \\ \sum_{v=1}^{H} \beta_{t+1}(v) q_{sv}(t) f(\mathbf{X}_{t+1}|S_{t+1} = v, \Theta), & t = 1, \cdots, T-1 \end{cases} \tag{7}$$

$$\begin{aligned} \xi_t(v,s) &:= P(S_t = v, S_{t+1} = s|\mathbf{X}_1, \cdots, \mathbf{X_T}, \Theta) \\ &= \frac{\alpha_t(v) q_{vs}(t) f(\mathbf{X}_{t+1}|S_{t+1} = s, \Theta) \beta_{t+1}(s)}{P(\mathbf{X}_1, \cdots, \mathbf{X}_T|\Theta)} \\ &\propto \alpha_t(v) q_{vs}(t) f(\mathbf{X}_{t+1}|S_{t+1} = s, \Theta) \beta_{t+1}(s) \end{aligned} \tag{8}$$

$$\gamma_t(s) := P(S_t = s|\mathbf{X}_1, \cdots, \mathbf{X}_T, \Theta) = \frac{\alpha_t(s)\beta_t(s)}{\sum_{v=1}^{H} \alpha_t(v)\beta_t(v)} \tag{9}$$

The joint log likelihood of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T$ and $S_1, \cdots, S_T$ may be written as

$$\begin{aligned} \log P(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T, S_1, \cdots, S_T|\Theta) &= \log P(S_1) + \sum_{t=1}^{T-1} \log q_{S_t, S_{t+1}}(t) \\ &+ \sum_{t=1}^{T} \log f(\boldsymbol{X}_t|S_t, \mu_{S_t}, \Sigma_{S_t}) \end{aligned} \tag{10}$$

Given old parameters Θ, the EM algorithm first computes the conditional expectation (E-step) of the joint log likelihood: $EQ(\Theta, \widehat{\Theta}) = E[\log P(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T, S_1, \cdots, S_T|\widehat{\Theta})|\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T, \Theta]$, then it looks for a $\widehat{\Theta}$ which maximizes $EQ(\Theta, \widehat{\Theta})$ (M-step) and takes the maximizer as the updated parameters

Θ. In view of Eq. (6), (7), (8), (9), we have

$$\begin{aligned} EQ(\Theta, \widehat{\Theta}) &= \sum_{s=1}^{H} \gamma_1(s) \log(\widehat{\pi}_s) + \sum_{v,s=1}^{H} \sum_{t=1}^{T-1} \xi_t(v,s) \log \widehat{q}_{vs}(t) \\ &+ \sum_{s=1}^{H} \sum_{t=1}^{T} \gamma_t(s) \log f(\boldsymbol{X}_t | \widehat{\mu}_s, \widehat{\Sigma}_s) \end{aligned} \tag{11}$$

In Eq. (11), the $\gamma_t(s)$ and $\xi_t(v,s)$ are obtained by Eq. (6), (7), (8), (9) with the old parameters Θ. The M-step is an optimization problem with probability constraints $\sum_{s=1}^{H} \widehat{\pi}_s = 1$. The updating formula for π follows trivially by introducing Lagrange multipliers.

$$\widehat{\pi}_s = \gamma_1(s) \tag{12}$$

The $\widehat{\mu}, \widehat{\Sigma}$ is the maximizer of the third term of $EQ(\Theta, \widehat{\Theta})$, which is a weighted least-square regression problem. So the updating formula is readily accessible:

$$\widehat{\mu}_s = \sum_{t=1}^{T} \gamma_t(s) \boldsymbol{X}_t / \sum_{t=1}^{T} \gamma_t(s), \quad \widehat{\Sigma}_s = \sum_{t=1}^{T} \gamma_t(s) (\boldsymbol{X}_t - \widehat{\mu}_s)(\boldsymbol{X}_t - \widehat{\mu}_s)^t / \sum_{t=1}^{T} \gamma_t(s) \tag{13}$$

Unfortunately, the transition probability parameters λ, $\boldsymbol{q}$ for $P(S_{t+1}|S_t, \boldsymbol{X}_t)$ do not have closed-form updating formulas, contrary to the regular HMM case. Instead, they are obtained by maximizing the second term of $EQ(\Theta, \widehat{\Theta})$:

$$\widehat{\lambda}, \widehat{\boldsymbol{q}} = \arg\max_{\widehat{\lambda}, \widehat{\boldsymbol{q}}} \sum_{v,s=1}^{H} \sum_{t=1}^{T-1} \xi_t(v,s) \log \widehat{q}_{vs}(t) \tag{14}$$

To this end, we adopt the BFGS algorithm, Nocedal and Wright (1999), by providing

the gradients of our objective function. In fact, referring to Eq. (3), we have

$$\begin{aligned}\partial \log \widehat{q}_{vs}(t)/\partial \lambda_{s'v'} &= \partial[\lambda_{sv} + \mathbf{q}_s^t \mathbf{X}_t - \log \sum_{\xi=1}^{H} \exp(\lambda_{\xi v} + \mathbf{q}_\xi^t \mathbf{X}_t)]/\partial \lambda_{s'v'} \\ &= [\delta_{s's} - \widehat{q}_{vs'}(t)]\delta_{v'v} \end{aligned} \tag{15}$$

$$\partial \log \widehat{q}_{vs}(t)/ \bigtriangledown \boldsymbol{q}_{s'} = [\delta_{s's} - \widehat{q}_{vs'}(t)]\boldsymbol{X}_t \tag{16}$$

where $\delta_{s's} = 1$ if $s' = s$ and zero otherwise. All the remains is to compute the gradient of $\sum_{v,s=1}^{H} \sum_{t=1}^{T-1} \xi_t(v,s) \log \widehat{q}_{vs}(t)$. This speeds up the convergence to a great extent compared with derivative-free algorithms such as simplex method, Press *et al.* (1992).

If our training sample has M *independent* sequences that are realizations on the same HMM model with same parameters, the Eq. (11) takes the form

$$\begin{aligned} EQ(\Theta, \widehat{\Theta}) &:= \sum_{e=1}^{M} E[\log P(\boldsymbol{X}_1^e, \cdots, \boldsymbol{X}_{T_e}^e, S_1^e, \cdots, S_{T_e}^e | \widehat{\Theta}) | \boldsymbol{X}_1^e, \cdots, \boldsymbol{X}_{T_e}^e, \Theta] \\ &= \sum_{e=1}^{M} \{ \sum_{s=1}^{H} \gamma_1^e(s) \log(\widehat{\pi}_s) + \sum_{v,s=1}^{H} \sum_{t=1}^{T_e - 1} \xi_t^e(v,s) \log \widehat{q}_{vs}^e(t) \\ &+ \sum_{s=1}^{H} \sum_{t=1}^{T_e} \gamma_t^e(s) \log f(\boldsymbol{X}_t^e | \widehat{\mu}_s, \widehat{\Sigma}_s) \} \end{aligned} \tag{17}$$

where $\xi_t^e(v,s), \gamma_t^e(s)$ are defined by Eq. (8), (9) for the eth sequence in the training sample. The constraints on $\widehat{\pi}, \widehat{q}_{vs}$ can be implemented by Lagrange multiplier. The M-step updating formula is, in analogy with Eq. (12) and (13),

$$\widehat{\pi}_s = \sum_{e=1}^{M} \gamma_1^e(s) / \sum_{e=1}^{M} \sum_{v=1}^{H} \gamma_1^e(v), \quad \widehat{\lambda}, \widehat{\boldsymbol{q}} = \arg\max_{\widehat{\lambda}, \widehat{\boldsymbol{q}}} \sum_{e=1}^{M} \sum_{t=1}^{T_e - 1} \xi_t^e(v,s) \log \widehat{q}_{vs}^e(t) \tag{18}$$

$$
\begin{aligned}
\widehat{\mu}_s &= \sum_{e=1}^{M}\sum_{t=1}^{T_e}\gamma_t^e(s)\boldsymbol{X}_t^e / \sum_{e=1}^{M}\sum_{t=1}^{T_e}\gamma_t^e(s), \\
\widehat{\Sigma}_s &= \sum_{e=1}^{M}\sum_{t=1}^{T_e}\gamma_t^e(s)(\boldsymbol{X}_t^e - \widehat{\mu}_s)(\boldsymbol{X}_t^e - \widehat{\mu}_s)^t / \sum_{e=1}^{M}\sum_{t=1}^{T_e}\gamma_t^e(s)
\end{aligned} \tag{19}
$$

EM algorithm updates the parameters iteratively until convergence as judged by some preset criterion. In this work, the criterion is $\|\Theta - \widehat{\Theta}\|_1 < 0.00001$ where $\|.\|_1$ is the $\mathcal{L}_1$-norm.

3.3 Classification of latent state sequence

Given a data set $\{\widetilde{\boldsymbol{X}}_1, \cdots, \widetilde{\boldsymbol{X}}_{T'}\}$ which is presumably generated from the same HMM, the goal here is to use the estimated parameters of HMM, still denoted as Θ, to classify different time points by assigning state labels to them. If $\{\widetilde{\boldsymbol{X}}_1, \cdots, \widetilde{\boldsymbol{X}}_{T'}\}$ is the same as the training data, then it is an in-sample classification, otherwise it's out-of-sample classification. Since the two kinds of classification have the same methodology, we will focus on in-sample classification. According to the Bayes rule under 0-1 loss, the classification problem reduces to find $s_1, \ldots, s_t$ such that

$$
\begin{aligned}
(s_1, \ldots, s_T) &= \arg\max_{S_1,\ldots,S_T} P(S_1, \ldots, S_T | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, \Theta) \\
&= \arg\max_{S_1,\ldots,S_T} [P(S_1, \ldots, S_T | \pi, \lambda, Q) P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T | S_1, \ldots, S_T, \mu, \Sigma)]
\end{aligned} \tag{20}
$$

Since the state space of S_t is finite, the solution of Eq. (20) can be obtained by an exhaustive search, which is very inefficient. We turn to the Viterbi algorithm,

Viterbi (1967), an efficient procedure to search the whole set of all possible sequence of $\{S_1, \cdots, S_T\}$. The protocol of Viterbi algorithm consists of Eq. (21) $-$ (25).

For each specific value s, let $M_1(s)$ be the joint probability of the event that the initial observation is X_1 and $S_1 = s$. Then

$$M_1(s) = P(S_1 = s)P(\boldsymbol{X}_1|S_1 = s) = \pi_s P(\boldsymbol{X}_1|S_1 = s) \tag{21}$$

For $t > 1$, let $M_t(s)$ be the maximum joint posterior probability over all possible values of $S_1, \ldots, S_{t-1}$ with $S_t = s$, i.e.

$$M_t(s) = \max_{S_1,\ldots,S_{t-1}} P(S_1, \ldots, S_{t-1}, S_t = s, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t). \tag{22}$$

Then

$$\begin{aligned} M_t(s) &= \max_{S_1,\ldots,S_{t-1}} [P(S_1, \ldots, S_{t-1}, S_t = s|\lambda, Q)P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t|S_1, \ldots, S_{t-1}, S_t = s, \mu, \Sigma)] \\ &= \max_v [\underbrace{P(\boldsymbol{X}_t, S_t = s|\boldsymbol{X}_{t-1}, S_{t-1} = v)}_{b_t(v,s)} M_{t-1}(v)] \\ &= \max_v [b_t(v, s)M_{t-1}(v)]. \end{aligned} \tag{23}$$

For $n > 1$, let $\Psi_n(s)$ be the maximizer of the right hand of Eq. (23), i.e.

$$\Psi_n(s) = \arg\max_v [b_n(v, s)M_{n-1}(v)]. \tag{24}$$

Then the MAP estimate of the states can be recursively obtained as

$$s_T = \arg\max_s M_T(s),\ s_{t-1} = \Psi_t(s_t),\ t = T, \cdots, 2 \tag{25}$$

Notice that our NHMM model $b_t(v, s) = P(\boldsymbol{X}_t, S_t = s|\boldsymbol{X}_{t-1}, S_{t-1} = v) = q_{vs}(t - 1)f(\boldsymbol{X}_t|\mu_s, \Sigma_s)$ but HMM has $q_{vs}f(\boldsymbol{X}_t|\mu_s, \Sigma_s)$.

3.4 Gaussian mixture model based clustering

If we don't consider the serial correlation of $\{\boldsymbol{X}_t\}$, we may treat $\boldsymbol{X}_t$ as independent sample from a distribution of Gaussian mixture model (GMM) $f(\boldsymbol{X}) = \sum_{s=1}^{H} \pi_s f_s(\boldsymbol{X}|\Theta_s)$ with H components and π_s is the mixing proportion of component s. Each component probability distribution $f_s(\boldsymbol{X}|\Theta_s)$ describes the distribution of $\boldsymbol{X}_t$ given its latent state $S_t = s$. With known π_s and $\Theta = \{\Theta_s | s = 1, \ldots, H\}$, we may classify data into meaningful groupings by simple rule of $\arg\max_s \pi_s f_s(\boldsymbol{X}_t|\Theta_s)$. In this work each group corresponds to a particular latent state. This classification approach does not consider the serial correlation, in other words, $\{\boldsymbol{X}_t\}$ are treated as independent sample, which is certainly inadequate from modelling perspective. However, if classification is our ultimate goal, this inadequacy might not be a serious issue provided that the component probability distributions $f_s(\boldsymbol{X}|\Theta_s)$ are sufficiently distinguishable in the sense of negligible probability of overlapping region between any pair of $f_s(\boldsymbol{X}|\Theta_s)$ and $f_v(\boldsymbol{X}|\Theta_v)$. For instance, two normal distributions $N(0, 0.04)$ and $N(3, 0.25)$ are well separated whereas $N(0, 0.04)$ and $N(0.5, 0.25)$ are not.

The parameters π_s, Θ are estimated from training data by maximum likelihood estimate via EM algorithm. To see why EM algorithm is a natural vehicle for this problem, we define, for each time point t, a latent indicator vector

$\boldsymbol{I}_t = (I_{1,t}, \ldots, I_{H,t}) \in \mathcal{R}^H$ where $I_{i,t} = 1$ if $S_t = i$ and 0 otherwise. Assuming each $\boldsymbol{I}_t$ is *iid* according to a multinomial distribution of one draw with probability $\pi_1, \ldots, \pi_H$, the joint log likelihood is

$$l(\pi, \Theta | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, \boldsymbol{I}_1, \ldots, \boldsymbol{I}_T) = \sum_{t=1}^{T} \sum_{s=1}^{H} I_{s,t} \log[\pi_s f_s(\boldsymbol{X}_t | \Theta_s)] \tag{26}$$

Notice $\widetilde{I}_{s,t} := E(I_{s,t} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, \Theta, \pi) = \pi_s f_s(\boldsymbol{X}_t | \Theta_s) / \sum_{v=1}^{H} \pi_v f_v(\boldsymbol{X}_t | \Theta_v)$. So in analogy with Eq. (11), the E-step is given by

$$\begin{aligned} EQ(\Theta, \pi; \widehat{\Theta}, \widehat{\pi}) &= \sum_{t=1}^{T} \sum_{s=1}^{H} \widetilde{I}_{s,t} \log[\pi_s f_s(\boldsymbol{X}_t | \widehat{\Theta}_s)] \\ &= \sum_{s=1}^{H} [\underbrace{\sum_{t=1}^{T} \widetilde{I}_{s,t}}_{n_s}] \log(\widehat{\pi}_s) + \sum_{t=1}^{T} \sum_{s=1}^{H} \widetilde{I}_{s,t} \log[f_s(\boldsymbol{X}_t | \widehat{\Theta}_s)] \end{aligned} \tag{27}$$

Assuming $f_s(\boldsymbol{X} | \Theta_s) \sim N(\mu_s, \Sigma_s)$, we have closed-form expressions from M-step obtained by Lagrange multiplier, namely

$$\widehat{\pi}_s = n_s / T, \quad \widehat{\mu}_s = \sum_{t=1}^{T} \widetilde{I}_{s,t} \boldsymbol{X}_t / n_s, \quad \widehat{\Sigma}_s = \sum_{t=1}^{T} \widetilde{I}_{s,t} (\boldsymbol{X}_t - \widehat{\mu}_s)(\boldsymbol{X}_t - \widehat{\mu}_s)^t / n_s \tag{28}$$

The updating formulas for $\widehat{\Sigma}_s$ in Eq. (13) and (28) assume a full covariance matrix structure. For a different parametrization (diagonal, spherical, etc), the updating formula differs, see Celeux and Govaert (1995) for more details. Observe that $\widetilde{I}_{s,t} = P(S_t = s | \boldsymbol{X}_1, \ldots \boldsymbol{X}_T)$ is the estimated posterior probability distribution of S_t , so the classification of hidden state at time t is simply $\arg\max_s \widetilde{I}_{s,t}$.

4 Simulation Study

We present a simulation study to investigate the effectiveness of various classification approaches. The hidden Markov chain is generated according to $P(S_1) \sim \pi$ with the probability transition matrix Q. The observations $P(\boldsymbol{X}_t|S_t = s) \sim N(\mu_s, \Sigma_s), s \in \{1, 2\}$. The parameters are in Eq. (31) and (32). These parameters are in fact estimated from a training sample. We will refer to the simulation result in our later discussions.

The length of each simulated data sequence is 1200. The simulations are repeated for 1000 times. For each simulated sequence, we fit GMM and HMM respectively by EM algorithm to obtain parameter estimates. The classification is performed with estimated parameters. We also did the classification by Viterbi algorithm with true parameters to estimate the best possible classification accuracy.

In Table 1, the mean and standard deviations are computed from 1000 replicates. From Table 1, we make the following observations. First, the classification accuracy by Viterbi algorithm with the estimated HMM model parameters is as good as that with true parameters. The nearly perfect accuracy (over 99%) demonstrates the effectiveness of Viterbi algorithm in classifying the latent states. Moreover, the estimated maximum log likelihood of HMM by EM algorithm is statistically equal to the true log likelihood, which proves EM algorithm reliable for maximizing likelihood. Second, the GMM is clearly inadequate as its maximum log likelihood is far less than the true log

likelihood. Fig. 8 plots the estimated probability density of these true log likelihood from 1000 replicates. The density is estimated by kernel method with constant bandwidth. The mean value, -3759.75, of the maximum GMM log likelihood reported in Table 1 near the location of the vertical line labeled "GMM" in this plot, which is very extreme compared with the density curve of true log likelihood. Furthermore, the autocorrelation function of $\boldsymbol{X}_t$ reveals strong serial correlations (Fig. 6), whereas GMM treat $\boldsymbol{X}_t$ as independent samples.

Third, even though GMM is inadequate, its classification accuracy competes very well with that of the correct model, HMM. This can be explained by Fig. 7. Clearly the distributions of observations $\boldsymbol{X}_t$ from two groups $S = 1$ and $S = 2$ are well separated. So the serial correlation of $\boldsymbol{X}_t$ brings little advantage to classification. For each simulated sequence we compute its log likelihood with true parameters.

Table 1: Comparison of classification performance of Gaussian mixture model (GMM) and HMM when the data is generated from a HMM with conditional Gaussian observations.

	GMM(%)	V.TruePar(%)	V.EMPar(%)	GMM.L	True.L	HMM.L
Mean	94.40	99.17	99.16	-3759.75	-3373.58	-3358.48
SD	0.83	0.31	0.31	113.44	87.28	87.25

Note: The columns are, from left to right, classification accuracy by GMM, Viterbi algorithm with true HMM model parameters, Viterbi algorithm with estimated HMM parameters by EM algorithm, estimated maximum log likelihood function of GMM, log likelihood of HMM with true parameters and estimated log likelihood of HMM.

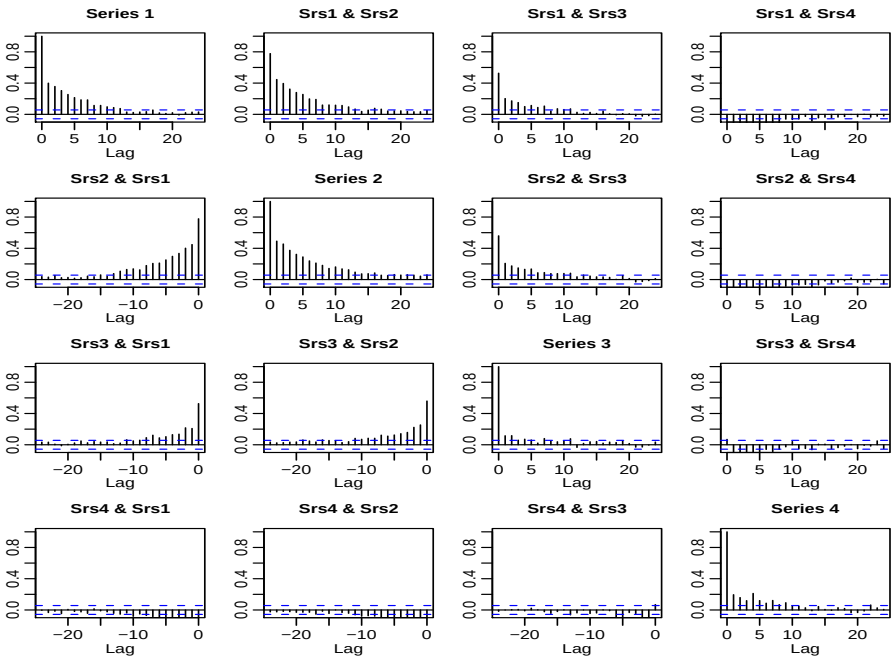

Figure 6: Autocorrelation function of $\{\boldsymbol{X}_t\}_{t=1,\ldots,1200}$ simulated from the HMM model.

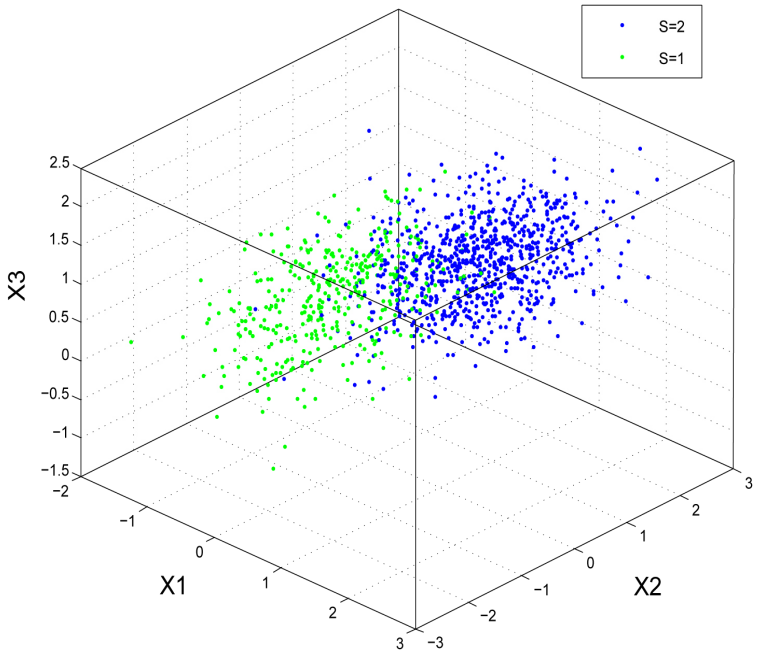

Figure 7: 3-D scatterplot of $\{\boldsymbol{X}_t\}_{t=1,\ldots,1200}$ simulated from the HMM model with parameters in Eq. (31) and (32). Only the first three components of $\boldsymbol{X}$ are plotted.

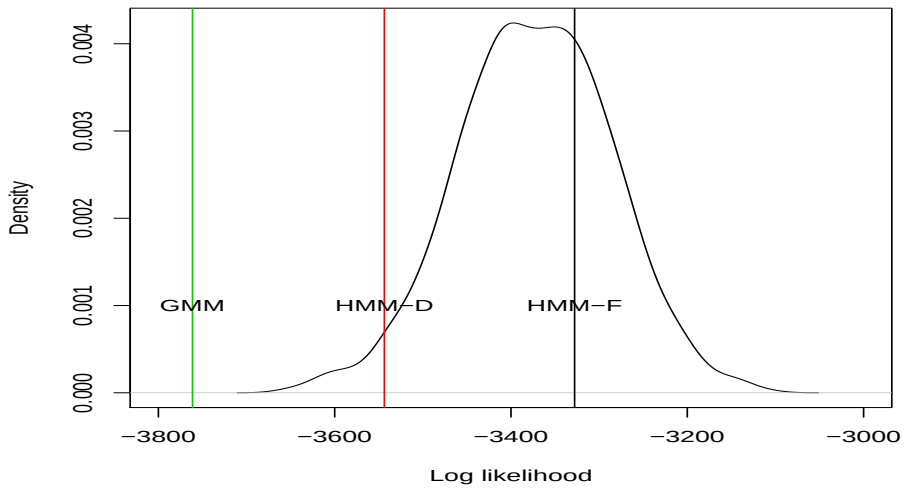

Figure 8: Empirical probability density of the true log likelihood of sequence $\{\boldsymbol{X}_t, t = 1, \ldots, 1200\}$ generated from HMM-F with parameters in Eq. (31), (32), 1000 replicates.

Note: The vertical lines locate the log likelihood of sample D estimated by GMM, HMM-D and HMM-F reported in Table 3, and the log likelihood of NHMM-F is -3357.91 which is very close to HMM-F, -3327.85.

5 Application to real data

The original EEG signal is represented by the time series $\{\boldsymbol{X}_t\}$ of its multi-band scores of PSD at a rate of 1 score per 0.6 second. We choose five disjoint training samples (A, B, C, D, E) from this time series. Each of the sample is a continuous block of $\{\boldsymbol{X}_t\}$ with no elements equal to the threshold A. This reduces the effect of movement artifact. The time series lengths are 1000, 900, 400, 1200 and 1000, corresponding to a time period of 10, 9, 4, 12 and 10 minutes, respectively.

5.1 Model fitting and classification

For each training sample, we fit the model GMM of two components and HMM with two latent states (REM/NREM). Then an in-sample classification is performed. For GMM, the best model is selected according to the Bayesian information criteria (BIC), see Fraley and Raftery (2002) for details. For HMM and NHMM, we consider two different parameterizations of the covariance matrices, HMM-D with diagonal covariance matrix $\Sigma_s = \mathrm{Cov}(\boldsymbol{X}_t|S_t = s)$ and HMM-F with a full covariance matrix Σ_s. Both HMM and GMM algorithms converge quite fast (less than a minute) for all training samples. Estimation of parameters of NHMM-F by EM algorithm is slightly slower (typically converges within 3 minutes) due to Eq. (14).

We report the results of model fitting with sample D. For the other training

data we include related results in various tables. Eq.(29)-(30),(31)-(32), (33)-(34) summarize the estimated parameters of GMM, HMM-F and NHMM-F, respectively. The subscripts 1 and 2 correspond to REM and NREM. The estimated parameters of HMM-D and NHMM-D are not reported due to space restriction. The estimated $\widehat{\pi} = P(S_1)$ for HMM has zero probability for state 2. This phenomenon of degenerated initial probability is typical when we fit a HMM to a single sequence. If the training sample has more than one sequence, then the estimated initial probability usually behaves normally. The agreement between the estimated μ, Σ by GMM and HMM-F is obvious. The maximum log likelihood of sample D obtained by GMM and HMM-F is -3761.32, -3327.85, respectively. The fact that HMM-F captures the serial correlations of $\{\boldsymbol{X}_t\}$ leads to a big increase of the log likelihood. Therefore, HMM-F is preferred by BIC, $-2\log likelihood + n\log(T)$, where n is the number of parameters.

Table 2 reports the classification accuracy on sample D and an out-of-sample test. For sample D, the overall in-sample accuracy of GMM is 81.1% with state specific accuracy as 58.3% (REM) and 94.6% (NREM). HMM-F achieves an accuracy of 85.25% (overall), 72.4% (REM) and 92.8% (NREM). The accuracy of NHMM-F is 86.08% (overall), 71.52% (REM) and 94.69% (NREM). The out-of-sample classification is performed on sample data E with estimated parameters from sample D. For GMM, the state of $\boldsymbol{Y}_t$ is classified by maximum posterior probability

$\arg\max_s P(S_t = s|\boldsymbol{Y}_t, \widehat{\mu}_s, \widehat{\Sigma}_s) = \arg\max_s \widehat{\pi}_s f(\boldsymbol{Y}_t|\widehat{\mu}_s, \widehat{\Sigma}_s)$. For HMM-F and NHMM-F, the classification is performed according to the Viterbi algorithm, see Eq. (25). GMM correctly classified 73.7% of the latent states in sample E with state specific accuracy of 65.7% (REM) and 84.3% (NREM). HMM-F has an overall accuracy of 81.0% with 80.3% of REM states and 81.9% of NREM states being correctly labeled. NHMM-F obtains accuracy of 78.60% (overall), 72.54% (REM) and 86.57% (NREM).

$$(\widehat{\pi}_1, \widehat{\pi}_2) = (0.26, 0.74), \qquad [\widehat{\mu}_1, \widehat{\mu}_2] = \begin{pmatrix} -0.34 & 0.76 \\ -0.46 & 0.81 \\ 0.64 & 0.97 \\ 1.54 & 0.93 \end{pmatrix} \tag{29}$$

$$\widehat{\Sigma}_1 = \begin{pmatrix} 0.33 & 0.21 & 0.23 & 0.13 \\ 0.21 & 0.34 & 0.23 & 0.09 \\ 0.23 & 0.23 & 0.46 & 0.14 \\ 0.13 & 0.09 & 0.14 & 0.48 \end{pmatrix}, \quad \widehat{\Sigma}_2 = \begin{pmatrix} 0.42 & 0.23 & 0.16 & 0.01 \\ 0.23 & 0.34 & 0.17 & 0.02 \\ 0.16 & 0.17 & 0.26 & 0.03 \\ 0.01 & 0.02 & 0.03 & 0.13 \end{pmatrix} \tag{30}$$

$$(\widehat{\pi}_1, \widehat{\pi}_2) = (1, 0), \ \widehat{Q} = \begin{pmatrix} 0.94 & 0.06 \\ 0.03 & 0.97 \end{pmatrix}, \ [\widehat{\mu}_1, \widehat{\mu}_2] = \begin{pmatrix} -0.30 & 0.85 \\ -0.40 & 0.90 \\ 0.56 & 1.04 \\ 1.45 & 0.92 \end{pmatrix} \tag{31}$$

$$\widehat{\Sigma}_1 = \begin{pmatrix} 0.31 & 0.17 & 0.16 & 0.10 \\ 0.17 & 0.32 & 0.16 & 0.06 \\ 0.16 & 0.16 & 0.39 & 0.15 \\ 0.10 & 0.06 & 0.15 & 0.43 \end{pmatrix}, \ \widehat{\Sigma}_2 = \begin{pmatrix} 0.36 & 0.16 & 0.11 & 0.03 \\ 0.16 & 0.27 & 0.12 & 0.04 \\ 0.11 & 0.12 & 0.22 & 0.04 \\ 0.03 & 0.04 & 0.04 & 0.13 \end{pmatrix} \tag{32}$$

$$(\widehat{\pi}_1, \widehat{\pi}_2) = (1, 0), \ \widehat{\lambda} = \begin{pmatrix} 0 & 0 \\ -2.20 & 2.20 \end{pmatrix}, \ \widehat{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.94 & 0.79 & 0.22 & 0.76 \end{pmatrix} \tag{33}$$

$$\begin{aligned} [\widehat{\mu}_1, \widehat{\mu}_2] &= \begin{pmatrix} -0.34 & 0.83 \\ -0.43 & 0.88 \\ 0.57 & 1.02 \\ 1.48 & 0.92 \end{pmatrix}, \widehat{\Sigma}_1 = \begin{pmatrix} 0.08 & 0.03 & 0.03 & 0.01 \\ 0.03 & 0.09 & 0.03 & 0.01 \\ 0.03 & 0.03 & 0.16 & 0.02 \\ 0.01 & 0.01 & 0.02 & 0.19 \end{pmatrix}, \\ \widehat{\Sigma}_2 &= \begin{pmatrix} 0.13 & 0.03 & 0.01 & 0.00 \\ 0.03 & 0.08 & 0.02 & 0.00 \\ 0.01 & 0.02 & 0.06 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.02 \end{pmatrix} \end{aligned} \tag{34}$$

Table 3 summarizes the classification performance on different training data by GMM, HMM-F, HMM-D, NHMM-D and NHMM-F, where HMM-D is the HMM

with a diagonal covariance matrix of $\mathrm{Cov}(\boldsymbol{X}_t|S_t)$ while HMM-F has full covariance matrices. As far as classification is concerned, HMM-F has slight edge over HMM-D and GMM except on sample B where GMM outperforms the others. The overall accuracy is around 80%.

Table 2: Classification result on training sample D and out-of-sample test on sample E.

		Training Data D		Test Data E	
Model		cREM	cNREM	cREM	cNREM
HMM-F	REM (%)	323 (72.42)	123	456 (80.28)	112
	NREM (%)	54	700 (92.84)	78	354 (81.94)
GMM	REM (%)	260 (58.30)	186	373 (65.67)	195
	NREM (%)	41	713 (94.56)	68	364 (84.26)
NHMM-F	REM (%)	319 (71.52)	127	412 (72.54)	156
	NREM (%)	40	714 (94.69)	58	374 (86.57)

Note: cREM and cNREM are the classified REM and NREM respectively. (%) stands for the percentage of correctly classification.

5.2 Discussion

We make several comments on the result above.

- HMM-F model consistently outperforms HMM-D across all training samples. In fact, one would expect this since HMM-D can be considered as a reduced model of HMM-F.

- The classification performances of NHMM-F and HMM-F are similar on single training samples, Table 3. Whereas on multiple training samples, NHMM-F

outperforms HMM-F on all of the five samples, Table 4. For the data from another bird (FEM1), NHMM-F has better performance than HMM-F in 3 of the total 4 samples, Table 5.

- GMM performs comparably with HMM-F and HMM-D. It seems that taking into account of the dependence of $\{\boldsymbol{X}_t\}$ over time does not lead to any big improvement in classification. This can be seen by the following observations. Let $\boldsymbol{X}_{-t}$ be the time series without $\boldsymbol{X}_t$, observe

$$\begin{aligned} \frac{P(S_t=s|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T)}{P(S_t=v|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T)} &= \frac{P(S_t=s,\boldsymbol{X}_t,\boldsymbol{X}_{-t})}{P(S_t=v,\boldsymbol{X}_t,\boldsymbol{X}_{-t})} \\ &= \frac{P(S_t=s,\boldsymbol{X}_{-t})P(\boldsymbol{X}_t|S_t=s,\boldsymbol{X}_{-t})}{P(S_t=v,\boldsymbol{X}_{-t})P(\boldsymbol{X}_t|S_t=v,\boldsymbol{X}_{-t})} \\ &= \frac{P(S_t=s|\boldsymbol{X}_{-t})}{P(S_t=v|\boldsymbol{X}_{-t})}\frac{P(\boldsymbol{X}_t|S_t=s)}{P(\boldsymbol{X}_t|S_t=v)} \end{aligned} \tag{35}$$

  In GMM, $\boldsymbol{X}_{-t}$ does not provide information on S_t therefore the odds ratio is completely determined by the very last fraction of Eq. (35). In other words, $\boldsymbol{X}_t$ from different time points are treated independently. HMM approach, however, accounts for the information of S_t carried by $\boldsymbol{X}_{-t}$, which is arguably at advantage. Nevertheless, the improvement of HMM, if any, could be marginal if the discriminating power based on the $P(\boldsymbol{X}_t|S_t)$ is already high, such as the fitted $\widehat{\mu}, \widehat{\Sigma}$ in Eq. (31) and (32).

- On the other hand, the classification accuracy reported in Table 3 is significantly less than that in Table 1 (85% vs. 99%). This suggests the HMM-F may be

inadequate. Moreover, Fig. 4 exhibits a long memory pattern as seen of slowly decaying ACF while the ACF in Fig. 6 has typical short memory behavior. So we conclude that HMM-F has not captured the time dependence of $\boldsymbol{X}_t$ adequately, this also explains why HMM-F has only small edge over GMM.

- As an assessment of goodness-of-fit, we perform a parametric bootstrap as follows. Having obtained the MLE from a training sample by EM algorithm, we generate 1000 sequences from the fitted model of same length as training sample, then compute the true log likelihood for each sequence. We compare the maximum log likelihood obtained from the training sample with the bootstrap samples of 1000 true log likelihood, which is reported by Fig. 8. Obviously, the maximum log likelihood of HMM-F on sample D falls in the high probability region, which suggests an adequate fit. However, this conflicts with the previous point in this discussion. One explanation is this parametric bootstrap is less powerful.

- Often it is of interest to assume the same model parameters across different training samples. In this regard, we fit models to the combined training sample of A-E, see Eq. (18)-(19) for training HMM on multiple samples. Fitting GMM on multiple samples is similar as on one sample. The result is reported in Table 4. The overall accuracy of HMM-F (80.03%) is almost identical to that

in Table 3. However, the accuracy of GMM (82.53%) improves dramatically and outperforms the others (NHMM-F, 82.33%).

- Adding more features into the model improves the performance as shown in Table 5, in which the upper half is achieved using the features of multi-band scores defined in Eq. (1), the lower half is obtained by including additional features of $\widetilde{X}_{ct} = \min\{Z_{ft}, \text{ for all frequnices } f \text{ in band } c\}$ subject to the same threshold. Besides the overall improvement of the accuracy, there is dramatic improvement on the last data sequence (IV), from about 60% to above 80%.

6 Future work and conclusion

In view of previous discussion, a plausible future research direction is to consider alternative model to capture the strong time dependence of the multi-band scores. This could be achieved by introducing certain autoregressive models for $P(\boldsymbol{X}_t|S_t, S_{t-1}, \boldsymbol{X}_{t-1})$. Second, comparison of model fitting and classification among different birds is also of interest to investigate any commonality. The last but not the least, different feature selections may affect classification power to a great extent, in this regard it is of absolute importance to consider alternative features besides the current one, for instance, we may add features other than $\widetilde{X}_{ct} = \min\{Z_{ft}, \text{ for all frequnices } f \text{ in band } c\}$ to selected features.

In this paper, we proposed a new representation, nonhomogeneous hidden Markov model, of EEG wave, multi-band scores of log power spectral densities, which is further modelled by readily accessible approaches, namely HMM and GMM, respectively. The classification accuracy ($75\% \sim 85\%$) is comparable to other more sophisticated and expensive approaches such as neural network.

Table 3: Classification performed on different training data

	Training sample	A	B	C	D	E
	Accuracy (%)	79.70	78.67	78.00	86.08	76.10
N-F	cNREM	559	559	141	714	373
	cREM	238	149	171	319	388
	Log likelihood	-2431.46	-2200.39	-1007.37	-3357.91	-2620.83
	Accuracy (%)	78.70	78.78	81.00	80.92	74.10
N-D	cNREM	558	564	137	581	332
	cREM	229	145	187	390	409
	Log Likelihood	-2520.78	-2301.99	-1024.58	-3485.21	-2662.91
	Accuracy (%)	80.10	77.67	80.50	85.25	75.50
H-F	cNREM	563	547	140	700	372
	cREM	238	152	182	323	383
	Log likelihood	−2417.26	−2184.94	−1007.08	−3327.85	−2632.25
	Accuracy (%)	78.20	77.44	80.25	80.83	74.20
H-D	cNREM	547	545	138	570	350
	cREM	235	152	183	400	392
	Log Likelihood	−2591.44	−2333.12	−1059.60	−3543.66	−2766.43
	Accuracy (%)	79.80	83.56	76.75	81.08	73.30
GMM	cNREM	581	624	135	713	357
	cREM	217	128	172	260	376
	Log likelihood	−2852.65	−2594.87	−1158.39	−3761.32	−2977.51
Total	NREM	752	745	159	754	432
	REM	248	155	241	446	568

Note: N-F and N-D are NHMM-F and NHMM-D respectively; H-F and H-D are HMM-F and HMM-D respectively. cNREM is the number of correctly classified NREM states, similarly for cREM. Accuracy is the percentage of correctly labelled states.

Table 4: Classification performed on combined multiple training samples A-E assuming common model parameters.

	Training sample	A	B	C	D	E
	Accuracy (%)	83.70	87.44	77.50	84.17	76.10
N-F	cNREM	625	659	86	646	275
	cREM	212	128	224	364	486
	Log likelihood	−2496.10	−2343.28	−1146.82	−3367.41	−2791.90
	Accuracy (%)	79.10	84.00	74.75	80.33	73.40
N-D	cNREM	588	626	80	605	257
	cREM	203	130	219	359	477
	Log Likelihood	-2643.84	-2484.88	-1247.95	-3576.10	-2908.24
	Accuracy (%)	82.30	85.00	72.50	81.83	74.30
H-F	cNREM	596	634	65	595	232
	cREM	227	131	225	387	511
	Log likelihood	−2481.79	−2337.27	−1161.26	−3374.64	−2814.99
	Accuracy (%)	82.00	83.89	75.50	80.83	73.00
H-D	cNREM	592	621	73	580	231
	cREM	228	134	229	390	499
	Log likelihood	−2647.40	−2488.87	−1252.64	−3589.18	−3000.13
	Accuracy (%)	84.50	88.11	80.75	83.50	75.10
GMM	cNREM	661	673	104	670	311
	cREM	184	120	219	332	440
	Log likelihood	−2925.47	−2759.34	−1326.39	−3839.91	−3195.63
Total	NREM	752	745	159	754	432
	REM	248	155	241	446	568

Note: N-F and N-D are NHMM-F and NHMM-D respectively; H-F and H-D are HMM-F and HMM-D respectively.

Table 5: Classification performed on single training data collected from another bird (FEM1).

	Training sample	I	II	III	IV
	Accuracy (%)	89.50	83.86	84.00	62.50
NHMM-F	cNREM	569	264	294	162
	cREM	147	323	168	88
	Log likelihood	−1840.91	−1692.42	−1126.06	−903.30
	Accuracy (%)	88.13	81.71	83.82	63.50
HMM-F	cNREM	554	250	289	162
	cREM	151	322	172	92
	Log likelihood	−1836.12	−1688.00	−1126.63	−904.57
	Accuracy (%)	84.34	80.14	79.45	62.50
GMM	cNREM	575	259	304	151
	cREM	100	302	133	99
	Log likelihood	−2082.56	−1947.03	−1353.27	−1028.17
Total states	NREM	604	337	322	274
	REM	196	363	228	126
	Accuracy (%)	87.12	81.86	79.45	86.50
NHMM-F	CNREM	553	234	312	261
	cREM	144	339	125	85
	Log likelihood	−4131.59	−3669.40	−2704.07	−2008.49
	Accuracy (%)	86.38	81.00	82.73	85.25
HMM-F	cNREM	541	251	298	255
	cREM	150	316	157	86
	Log likelihood	−4123.48	−3655.91	−2629.40	−2001.24
	Accuracy (%)	86.50	63.57	79.64	81.75
GMM	cNREM	582	98	302	261
	cREM	110	347	136	66
	Log likelihood	−4417.17	−4014.64	−2893.50	−2191.63

Note: The upper half is achieved using the features of multi-band scores defined in Eq. (1). The lower half is obtained by including additional features of $\widetilde{X}_{ct} = \min\{Z_{ft},$ for all frequnices f in band $c\}$

References

Amlander, C. and Ball, N. (1994). *Principles and Practice of Sleep Medicine. (Second edition)*, chap. 7. Philadelphia: Saunders, pp. 81–94.

Anderson, C., Devulapalli, S., and Stolz, E. (1995). Discriminating mental tasks using eeg represented by ar models. In *Proceedings of the 1995 IEEE Engineering in Medicine and Biology Annual Conference*. Montreal, Canada.

Andrieu, C. and Doucet, A. (2000). Simulated annealing for maximum *a posteriori* parameter estimation of Hidden Markov Models. *IEEE Transactions on Information Theory* **46** 994–1004.

Aserinsky, E. and Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science* **118** 273–274.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41** 164–171.

Brockwell, P. J. and Davis, R. A. (1991). *Time series. Theory and methods (Second edition)*. Springer-Verlag Inc.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28** 781–793.

Cohen, A., Flomen, F., and Drori, N. (1996). *Advances in Processing and Pattern Analysis of Biological Signals*, chap. 4. New York : Plenum Press, pp. 45–55.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: P22-37). *Journal of the Royal Statistical Society, Series B: Methodological* **39** 1–22.

Empson, J. (1989). *Sleep and dreaming*. London : Faber.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97** 611–631.

Gersch, W., F., M., J., Y., M.D., L., and J.A., M. (1979). Automatic classification of electroencephalograms:kullback-leibler nearest neighbor rules. *Science* **205** 193–195.

Gersch, W. (1996). *Advances in Processing and Pattern Analysis of Biological Signals*, chap. 1. New York : Plenum Press, pp. 1–19.

Haselsteiner, E. and Pfurtscheller, G. (2000). Using timedependent neural networks for EEG classification. *IEEE Transactions on Rehabilitation Engineering* **8** 457–463.

Kitagawa, G. (1987). Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* **82** 1032–1041.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models. (Second edition)*. Chapman Hall/CRC.

Mutapcic, A., Shimayama, T., and Flores, A. (2003). Automatic sleep stage classification using frequency analysis of EEG. In *Proceedings of XIX International Symposium on Information and Communication Technologies*. Sarajevo, Bosnia and Herzegovina.

Nick, T. and Konishi, M. (2001). Dynamic control of auditory activity during sleep: Correlation between song response and EEG. *Proceedings of the National Academy of Sciences* **98** 14012–14016.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer-Verlag Inc.

Penny, W. and Roberts, S. (1998). Gaussian observation hidden markov models for EEG analysis. *Technical Report* .

Penny, W. and Roberts, S. (1999). Dynamic models for nonstationary signal segmentation. *Computers and Biomedical Research* **32** 483–502.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C (Second edition)*. Cambridge University Press, pp. 408–412.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** 257–286.

Schmidt, D., Ball, N., and Amlaner, C. (1990). The characteristics and quantities of sleep in the Zebra finch *(genus taenopygia)*. *Sleep Research* **19** 111.

Sergejew, A. and Tsoi, A. (1996). *Advances in Processing and Pattern Analysis of Biological Signals*, chap. 3. New York : Plenum Press, pp. 33–44.

Thompson, D. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE* **70** 1055–1096.

Tobler, I. (2005). *Principles and Practice of Sleep Medicine. (Fifth edition)*, chap. 7. Philadelphia: Elsevier/Saunders, p. 77.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13** 260–267.