International Journal of Machine Learning and Cybernetics (2022) 13:3019–3029
https://doi.org/10.1007/s13042-022-01579-7

ORIGINAL ARTICLE

Unsupervised modeling and feature selection of sequential spherical data through nonparametric hidden Markov models

Wentao Fan[1,2] · **Wenjuan Hou**[3]

Received: 2 August 2021 / Accepted: 13 May 2022 / Published online: 6 June 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
As spherical data (i.e. L_2 normalized vectors) are often encountered in a variety of real-life applications (such as gesture recognition, gene expression analysis, etc.), sequential spherical data modeling has become an important research topic in recent years. Hidden Markov models (HMMs), as probabilistic graph models, have shown their effectiveness in modeling sequential data in previous research works. In this article, we propose a nonparametric hidden Markov model (NHMM) for modeling time series or sequential spherical data vectors. In our model, the emission distribution of each hidden state obeys a mixture of von Mises (VM) distributions which has better capability for modeling spherical data than other popular distributions (e.g. the Gaussian distribution). As we construct our NHMM by leveraging a Bayesian nonparametric model namely the Dirichlet process, the amount of hidden states and the number of mixture components for each state can be automatically adjusted according to observed data set. In addition, to handle high-dimensional data sets which may contain irrelevant or noisy features, feature selection, which is the process of selecting the "best" feature subset for describing the given data set, is adopted in our framework. In our case, an unsupervised localized feature selection method is incorporated with the developed NHMM, which results in a unified framework that can simultaneously perform data modeling and feature selection. Our model is learned by theoretically developing a convergence-guaranteed algorithm through variational Bayes. The advantages of our model are demonstrated by conducting experiments on both synthetic and real-world sequential data sets.

Keywords Hidden Markov model · Spherical data · Feature selection · Von Mises mixture · Dirichlet process · Variational Bayes

1 Introduction

With the rapid advancement in data acquisition technology, time series and sequential data modeling have become an important research topic in various domains, ranging from medical virus sequences and human genome sequences modeling [19], gesture recognition [31], abnormal behaviors detection [28] to text clustering [32]. One of the most powerful tools for modeling sequential data or time series is the hidden Markov model (HMM) [33, 34], which is a probability graphical model assuming that each data observation in a hidden state is generated based on a probability density (namely the emission distribution).

In the literature of HMMs, the Gaussian distribution or the Gaussian mixture model (GMM) are common choices as emission densities for HMMs to model continuous sequential observations [21, 41]. Nevertheless, a number of research works have shown that HMMs with other emission densities are better alternatives than Gaussian-based HMMs in various practical applications where data often possess non-Gaussian property (e.g. the distribution of data is normally not symmetric) [8, 11, 15, 30]. Among different types of data, the L_2 normalized data, also called *spherical data* as they are defined on a unit hypersphere [29], have drawn considerable attention as they are usually confronted in many real-world applications [25, 29], such as gene expression clustering, fMRI data analysis, text clustering, etc. Moreover, in a variety of applications, the

✉ Wentao Fan
fwt@hqu.edu.cn

[1] Department of Computer Science and Technology, Huaqiao University, Xiamen, China

[2] Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College (UIC), Zhuhai, China

[3] Instrumental Analysis Center, Huaqiao University, Fujian, Xiamen, China

L_2 normalization is commonly adopted as an essential preprocessing step to handel the issue of sparsity by restricting the data on a hypersphere. It also has been shown that the clustering performance can be improved for various models if L_2 normalization is applied during training [2]. In contrast with other distributions, a reasonable choice for modeling spherical data is through directional distributions, such as the von Mises (VM) distribution [7, 12, 29], the von Mises-Fisher (VMF) distribution [3, 29, 38], and the Watson distribution [14, 16, 36, 37]. Recently, an effective model has been proposed to model sequential spherical data based on HMM with VMF mixture models [15]. One limitation of this model is that the amount of hidden states for the HMM and the total number of VMF distributions for the VMF mixture model under each state are determined by treating the log-likelihood function as the model selection criterion. This method, however, demands high-computational resources and is time-consuming, since it has to implement the model learning algorithm multiple times with different numbers of hidden states and mixture components in order to obtain the optimal solution with the highest model selection scores. Another limitation of the HMM in [15] and many other exiting HMMs (such as [8, 11, 30], etc.) is that, they assume all features are equally important in data modeling. Nevertheless, this assumption is unsuccessful in real applications where high-dimensional data normally involve irrelevant features that may degrade the modeling performance. An effective solution to this problem is *feature selection* [17, 26], which is the process of selecting the "best" feature subset for describing the given data set. Recently, a variety of feature selection techniques [1, 10, 20, 40] have been developed and shown their effectiveness for handling high-dimensional data in different applications.

The goal of our work is to propose a novel nonparametric HMM (NHMM) for modeling sequential spherical observations. In our model, the emission distribution of each hidden state is distributed according to a VM mixture model which has better capability for modeling spherical data than other popular distributions (e.g. Gaussian distribution). Our NHMM is constructed by leveraging a Bayesian nonparametric framework namely as the *Dirichlet process* (DP) [39]. By applying the stick-breaking representation [35] of the DP in our NHMM, the amount of hidden states and the number of mixture components for each state can be automatically adjusted based on the observed data set. Moreover, to deal with high-dimensional data which may include irrelevant features, feature selection is adopted in our approach. Here, we formulate a unified framework which can simultaneously perform data modeling and feature selection by integrating an unsupervised localized feature selection method [13, 27, 42] in terms of *feature saliency* [24] with the proposed NHMM. The proposed model (namely VM-NHMM-Fs) is learned by theoretically developing a convergence-guaranteed algorithm based on variational Bayes (VB) [6, 22], which is a deterministic learning algorithm for approximating probability densities through optimization, and has been successfully applied in various Bayesian models. The advantages of our model are demonstrated by conducting experiments on both synthetic and real-world sequential data sets.

We summarize the contributions of our work as follows.

- A novel NHHM with VM mixture models as its emission densities is proposed for modeling sequential spherical data;
- The total number of hidden states and mixture components of our model are inferred automatically by leveraging the nonparametric stick-breaking DP;
- We integrate our model with a localized feature selection method which results in a unified framework for both data modeling and feature selection;
- A convergence-guaranteed algorithm based on VB inference is theoretically developed to learn the proposed model.

We organize the following parts of our paper as follows. We start by presenting the VM based NHMM with unsupervised localized feature selection in Sect. 2. We develop an effective approach based on VB inference in Sect. 3 to learn the proposed model. In Sect. 4, we report the experimental results using both synthetic and real-world data sets. Finally, we provide the conclusion in Sect. 5.

2 The nonparametric HMM with VM mixture model and localized feature selection

2.1 The VM mixture model with localized feature selection

A proper choice to model a D-dimensional spherical (i.e. L_2 normalized) vector $\boldsymbol{y} = \{y_d\}_{d=1}^D$ is the D-dimensional von Mises (VM) distribution [29]

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\mu}, \lambda) &= \prod_{d=1}^{D} \mathrm{VM}\big(\boldsymbol{x}_d|\boldsymbol{\mu}_d, \lambda_d\big) \\ &= \prod_{d=1}^{D} \frac{1}{2\pi I_0(\lambda_d)} \exp\left(\lambda_d \boldsymbol{\mu}_d^{\mathrm{T}} \boldsymbol{x}_d\right), \end{aligned} \tag{1}$$

where $\|\boldsymbol{y}\|_2 = 1$, $\boldsymbol{x}_d = (x_{d1}, x_{d2})$, and $x_{d1} = y_d$. It is noteworthy that x_{d2} is included in the vector $\boldsymbol{x}_d$ to attain the L_2 normalization of $\boldsymbol{x}_d$ (i.e., $\|\boldsymbol{x}_d\|_2 = 1$). $I_0(\cdot)$ represents the modified Bessel function of the first kind of order 0 [29]. The parameter $\boldsymbol{\mu} = \{\boldsymbol{\mu}_d\}_{d=1}^D$ indicates the mean direction, and

$\lambda = \{\lambda_d\}_{d=1}^{D}$ in (1) represents the concentration parameter, where $\boldsymbol{\mu}_d = (\mu_{d1}, \mu_{d2})$ and $\lambda_d \geq 0$.

A more flexible and powerful way to model the L_2 normalized D-dimensional vector $\boldsymbol{y}$ is though a mixture of K VM distributions as

$$p(\boldsymbol{y}|\boldsymbol{c}, \boldsymbol{\mu}, \lambda) = \sum_{k=1}^{K} c_k \prod_{d=1}^{D} \mathrm{VM}(\boldsymbol{x}_d|\boldsymbol{\mu}_{kd}, \lambda_{kd}), \tag{2}$$

where $\boldsymbol{c} = \{c_k\}_{k=1}^{K}$, $\sum_{k=1}^{K} c_k = 1$ represent mixing coefficients. As we may notice, all features in this VM mixture model (2) are equally treated. In practical applications, however, high-dimensional data often include noise or features that are irrelevant to the corresponding task. In our work, we solve this issue by adopting an unsupervised localized feature selection method [27]. The main idea is to assume that irrelevant features of the VM mixture model are distributed according to a common VM distribution that does not depend on class labels

$$p(y_d) = \mathrm{VM}\big(\boldsymbol{x}_d|\boldsymbol{\mu}_{kd}, \lambda_{kd}\big)^{z_{kd}} \mathrm{VM}\big(\boldsymbol{x}_d|\boldsymbol{\mu}'_{kd}, \lambda'_{kd}\big)^{1-z_{kd}}, \tag{3}$$

where the binary variable z_{kd} represents the feature relevancy in the kth component of the VM mixture model. If z_{kd} equals 0, it means that the dth feature associated with the kth VM density is irrelevant and is distributed as $\mathrm{VM}(\boldsymbol{x}_d|\boldsymbol{\mu}'_{kd}, \lambda'_{kd})$. When z_{kd} equals 1, it indicates that the dth feature is relevant and follows the VM distribution $\mathrm{VM}(\boldsymbol{x}_d|\boldsymbol{\mu}_{kd}, \lambda_{kd})$.

2.2 The VM-NHMM with localized feature selection

In this part, we propose a nonparametric HMM (NHMM) which is formulated through the stick-breaking representation of the DP. If an infinite VM mixture model (i.e. a VM mixture model with an infinite number of components) with localized feature selection is considered as the emission density of the NHMM with an infinite number of states, then the resulting VM-NHMM-Fs model can be defined with parameters $\boldsymbol{\Phi} = \{\boldsymbol{\pi}, A, C, \boldsymbol{\Theta}\}$, where $\boldsymbol{\pi} = \{\pi_i\}_{i}^{\infty}$ denotes the initial state probability matrix, $A = \{a_{ij}\}_{i,j}^{\infty,\infty}$ represents the state transition matrix, $C = \{c_{ik}\}_{i,k}^{\infty,\infty}$ is the mixing coefficient matrix, and $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \lambda, \boldsymbol{\mu}', \lambda'\}$ denotes the set of parameters that governs the VM densities with $\boldsymbol{\mu} = \{\boldsymbol{\mu}_{ikd}\}_{i,k,d}^{\infty,\infty,D}$, $\lambda = \{\lambda_{ikd}\}_{i,k,d}^{\infty,\infty,D}$, $\boldsymbol{\mu}' = \{\boldsymbol{\mu}'_{ikd}\}_{i,k,d}^{\infty,\infty,D}$, $\lambda' = \{\lambda'_{ikd}\}_{i,k,d}^{\infty,\infty,D}$.

Given a sequence of T observations $Y = \{\boldsymbol{y}_t\}_{t}^{T}$, where $\boldsymbol{y}_t = \{y_{td}\}_{td}^{TD}$ represents the feature vector at time t. $S = \{s_t\}_{t}^{T}$, where $s_t \in [1, \infty]$ indicates the hidden state associated with the tth observation. $L = \{l_t\}_{t}^{T}$, where $l_t \in [1, \infty]$ indicates from which component of the VM mixture model that the tth observation is generated. The latent variable $z = \{z_{tikd}\}_{t,i,k,d}^{T,\infty,\infty,D}$ represents the saliencies of different features in different components.

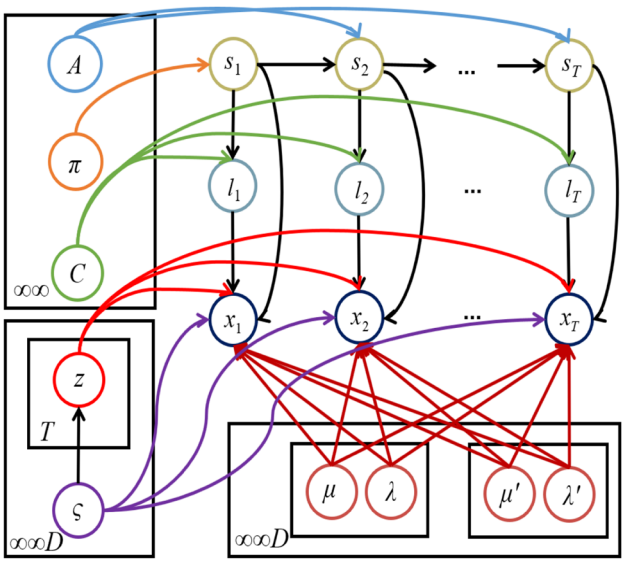

Fig. 1 Graphical model of the proposed VM-NHMM-Fs

The model diagram of VM-NHMM-FS is shown in Fig. 1, and the probability distribution of this model is given by

$$p(Y, S, L|\boldsymbol{z}, \boldsymbol{\Phi}) = \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} p(\boldsymbol{y}_t|\boldsymbol{\Theta}, \boldsymbol{z}_t)\right], \tag{4}$$

where $p(\boldsymbol{y}_t|\boldsymbol{\Theta}, \boldsymbol{z}_t)$ denotes the VM density with feature selection and can be represented by

$$p\big(\boldsymbol{y}_t|\boldsymbol{\Theta}, \boldsymbol{z}_t\big) = \prod_{d=1}^{D} \left[\mathrm{VM}\big(\boldsymbol{x}_{td}|\boldsymbol{\mu}_{s_t l_t d}, \lambda_{s_t l_t d}\big)^{z_{ts_t l_t d}} \mathrm{VM}\Big(\boldsymbol{x}_{td}|\boldsymbol{\mu}'_{s_t l_t d}, \lambda'_{s_t l_t d}\Big)^{1-z_{ts_t l_t d}}\right]. \tag{5}$$

Therefore, we can represent the likelihood of parameters $\boldsymbol{\Phi}$ for the data sequence Y as

$$p(Y|\boldsymbol{\Phi}) = \sum_{S,L} \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} p(\boldsymbol{y}_t|\boldsymbol{\Theta}, \boldsymbol{z}_t)\right]. \tag{6}$$

2.3 Priors over model parameters

Since the proposed VM-NHMM-Fs is a Bayesian model, each unknown variable is associated with a prior distribution. The prior probability of the indicator variable $\boldsymbol{z}$ is defined by

$$p(\boldsymbol{z}|\boldsymbol{\zeta}) = \prod_{t=1}^{T} \prod_{i=1}^{\infty} \prod_{k=1}^{\infty} \prod_{d=1}^{D} \zeta_{ikd}^{z_{tikd}} (1 - \zeta_{ikd})^{1-z_{tikd}}, \tag{7}$$

where ζ_{ikd} represents the feature saliency indicating whether the dth feature in the kth component associated with the ith state is relevant.

For parameters $\boldsymbol{\mu}$, λ, $\boldsymbol{\mu}'$, and λ' of the VM distributions, von Mises-Gamma priors are adopted

$$p(\boldsymbol{\mu}, \lambda) = \prod_{i=1}^{\infty}\prod_{k=1}^{\infty}\prod_{d=1}^{D} \mathrm{VM}\left(\boldsymbol{\mu}_{ikd}|\boldsymbol{m}_{ikd}, \beta_{ikd}\lambda_{ikd}\right)\mathcal{G}\left(\lambda_{ikd}|u_{ikd}, v_{ikd}\right), \tag{8}$$

$$p\left(\boldsymbol{\mu}', \lambda'\right) = \prod_{i=1}^{\infty}\prod_{k=1}^{\infty}\prod_{d=1}^{D} \mathrm{VM}\left(\boldsymbol{\mu}'_{ikd}|\boldsymbol{m}'_{ikd}, \beta'_{ikd}\lambda'_{ikd}\right)\mathcal{G}\left(\lambda'_{ikd}|u'_{ikd}, v'_{ikd}\right), \tag{9}$$

where $\boldsymbol{m}_{ikd} = \left(m_{ikd1}, m_{ikd2}\right)$ and $\boldsymbol{m}'_{ikd} = \left(m'_{ikd1}, m'_{ikd2}\right)$.

In our model, similar to [9], we adopt a nonparametric DP [39] as the prior over parameters $\boldsymbol{\pi}'$, A and C. According to the stick-breaking representation of the DP [35], π_i, c_{ik} and a_{ij} can be represented by

$$\pi_i = \pi'_i \prod_{n=1}^{i-1}(1-\pi'_n), \tag{10}$$

$$a_{ij} = a'_{ij} \prod_{n=1}^{j-1}\left(1-a'_{in}\right), \tag{11}$$

$$c_{ik} = c'_{ik} \prod_{n=1}^{k-1}\left(1-c'_{in}\right), \tag{12}$$

where $\boldsymbol{\pi}'$, A' and C' are distributed according to Beta distributions

$$p(\boldsymbol{\pi}') = \prod_{i=1}^{\infty}\mathrm{Beta}\left(1, \phi^{\pi}_i\right) = \prod_{i=1}^{\infty}\phi^{\pi}_i\left(1-\pi'_i\right)^{\phi^{\pi}_i-1}, \tag{13}$$

$$p(A') = \prod_{i=1}^{\infty}\prod_{j=1}^{\infty}\mathrm{Beta}(1, \phi^{A}_{ij}) = \prod_{i=1}^{\infty}\prod_{j=1}^{\infty}\phi^{A}_{ij}\left(1-a'_{ij}\right)^{\phi^{A}_{ij}-1}, \tag{14}$$

$$p(C') = \prod_{i=1}^{\infty}\prod_{k=1}^{\infty}\mathrm{Beta}(1, \phi^{C}_{ik}) = \prod_{i=1}^{\infty}\prod_{k=1}^{\infty}\phi^{C}_{ik}\left(1-c'_{ik}\right)^{\phi^{C}_{ik}-1}. \tag{15}$$

3 Model learning algorithm based on VB inference

In this section, we systematically develop an effective learning approach which is tailored for learning the proposed VM-NHMM-Fs through variational Bayes (VB). In our case, our goal is to discover a proper approximation $q(S, L, \boldsymbol{z}, \boldsymbol{\Phi})$ to the true posterior $p(S, L, \boldsymbol{z}, \boldsymbol{\Phi}|Y)$, where $\{S, L, \boldsymbol{z}, \boldsymbol{\Phi}\}$ denotes the set of latent and unknown variables in VM-NHMM-Fs as described previously. To obtain a tractable inference procedure, we apply the mean-field theory [4] as

$$q(\boldsymbol{z}, S, L, \boldsymbol{\Phi}) = q(\boldsymbol{z})q(S, L)q(\boldsymbol{\Phi}). \tag{16}$$

The approximations $q(\boldsymbol{z})$, $q(S, L)$ and $q(\boldsymbol{\Phi})$ (also known as variational posteriors) in VB inference can be found by maximizing the objective function, which is the evidence lower bound (ELBO) and is defined by

$$\begin{aligned}\mathrm{ELBO}(q) &= \int q(\boldsymbol{z}, S, L, \boldsymbol{\Phi})\ln\frac{p(Y, \boldsymbol{z}, S, L, \boldsymbol{\Phi})}{q(\boldsymbol{z}, S, L, \boldsymbol{\Phi})}\mathrm{d}\boldsymbol{z}\mathrm{d}S\mathrm{d}L\mathrm{d}\boldsymbol{\Phi} \\ &= \mathrm{ELBO}\left(q\left(\boldsymbol{\pi}'\right)\right) + \mathrm{ELBO}(q(A')) + \mathrm{ELBO}(q(C')) \\ &\quad + \mathrm{ELBO}(q(\boldsymbol{\Theta})) + \mathrm{ELBO}(q(\boldsymbol{z})) + \mathit{Constant}.\end{aligned} \tag{17}$$

In addition, the truncation technique [5] is adopted to truncate the variational posteriors at finite numbers of hidden states and mixture components at N and K, respectively as

$$\pi'_N = 1, \quad \sum_{i=1}^{N}\pi_i = 1, \quad \pi_i = 0 \ \text{ if } \ i > N, \tag{18}$$

$$a'_{iJ} = 1, \quad \sum_{j=1}^{N}a_{ij} = 1, \quad a_{ij} = 0 \ \text{ if } \ j > N, \tag{19}$$

$$c'_{iK} = 1, \quad \sum_{k=1}^{K}c_{ik} = 1, \quad c_{ik} = 0 \ \text{ if } \ k > K, \tag{20}$$

where N and K will be inferred automatically during the VB learning process.

3.1 Optimizing variational posteriors $q(\boldsymbol{\pi}')$, $q(\boldsymbol{C}')$ and $q(\boldsymbol{A}')$

The variational posteriors of the initial state probability matrix $q(\boldsymbol{\pi}')$, the state transition matrix $q(A')$, and the mixing coefficient matrix $q(C')$ can be optimized by maximizing the ELBO in (17) as

$$q(\boldsymbol{\pi}') = \prod_{i=1}^{N}\mathrm{Beta}\left(\pi'_i|\widehat{W}^{\pi}_i, \widetilde{W}^{\pi}_i\right), \tag{21}$$

$$q(A') = \prod_{i=1}^{N}\prod_{j=1}^{N}\mathrm{Beta}\left(a'_{ij}|\widehat{W}^{A}_{ij}, \widetilde{W}^{A}_{ij}\right), \tag{22}$$

$$q(C') = \prod_{i=1}^{N}\prod_{k=1}^{K}\mathrm{Beta}\left(c'_{ik}|\widehat{W}^{C}_{ik}, \widetilde{W}^{C}_{ik}\right), \tag{23}$$

where the hyperparameters of the above variational posteriors are given by

$$\widehat{W}_i^{\pi} = 1 + q(s_1 = i), \tag{24}$$

$$\widetilde{W}_i^{\pi} = \phi_i^{\pi} + \sum_{n=i+1}^{N} q(s_1 = n), \tag{25}$$

$$\widehat{W}_{ij}^{A} = 1 + \sum_{t=1}^{T-1} q\big(s_t = i, s_{t+1} = j\big), \tag{26}$$

$$\widetilde{W}_{ij}^{A} = \phi_{ij}^{A} + \sum_{t=1}^{T-1} \sum_{n=j+1}^{N} q\big(s_t = i, s_{t+1} = n\big), \tag{27}$$

$$\widehat{W}_{ik}^{C} = 1 + \sum_{t=1}^{T} q(s_t = i, l_t = k), \tag{28}$$

$$\widetilde{W}_{ik}^{C} = \phi_{ik}^{C} + \sum_{t=1}^{T} \sum_{n=k+1}^{K} q\big(s_t = i, l_t = n\big), \tag{29}$$

where the classic forward-backward algorithm as described in [34] is adopted to compute $q(s_1)$, $q(s_t, s_{t+1})$ and $q(s_t, l_t)$.

3.2 Optimizing variational posterior $q(\boldsymbol{z})$

By maximizing the ELBO with respect to the feature saliency indicator z, we can optimize the variational posterior $q(z)$ as

$$q(z) = \prod_{t=1}^{T} \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \varphi_{tikd}^{z_{tikd}} (1 - \varphi_{tikd})^{1 - z_{tikd}}, \tag{30}$$

where φ_{tikd} can be computed by

$$\varphi_{tikd} = \frac{\exp\left(\widetilde{\varphi}_{tikd}\right)}{\exp\left(\widetilde{\varphi}_{tikd}\right) + \exp\left(\widehat{\varphi}_{tikd}\right)}, \tag{31}$$

$$\begin{aligned} \widetilde{\varphi}_{tikd} = q(s_t = i, l_t = k) \Bigg[& \langle \lambda_{ikd} \boldsymbol{\mu}_{ikd}^{\mathrm{T}} \boldsymbol{x}_{td} \rangle \\ & - \left(\frac{\partial}{\partial \lambda_{ikd}} \ln I_0(\bar{\lambda}_{ikd}) \right) \left(\langle \lambda_{ikd} \rangle - \bar{\lambda}_{ikd}^{(t-1)} \right) \\ & - \ln I_0\big(\bar{\lambda}_{ikd}\big) \Bigg] + \ln \zeta_{ikd}, \end{aligned} \tag{32}$$

$$\begin{aligned} \widehat{\varphi}_{tikd} = q(s_t = i, l_t = k) \Bigg[& \langle \lambda'_{ikd} \boldsymbol{\mu}'^{\mathrm{T}}_{ikd} \boldsymbol{x}_{td} \rangle \\ & - \left(\frac{\partial}{\partial \lambda'_{ikd}} \ln I_0(\bar{\lambda}'_{ikd}) \right) \left(\langle \lambda'_{ikd} \rangle - \bar{\lambda}'_{ikd} \right) \\ & - \ln I_0\big(\bar{\lambda}'_{ikd}\big) \Bigg] + \ln(1 - \zeta_{ikd}), \end{aligned} \tag{33}$$

where $\langle \cdot \rangle$ denotes the calculation of expectation, $\frac{\partial}{\partial \lambda_{ikd}} \ln I_0\big(\bar{\lambda}_{ikd}\big) = \frac{I_1(\bar{\lambda}_{ikd})}{I_0(\bar{\lambda}_{ikd})}$ is obtained based on the property $I_0'(\kappa) = I_1(\kappa)$ of the modified Bessel function as discussed in [38].

The saliency of the dth feature in the kth component for the ith hidden state can be calculated by setting the derivative of ELBO with respect to ζ_{ikd} to zero as

$$\zeta_{ikd} = \frac{1}{T} \sum_{t=1}^{T} \langle z_{tikd} \rangle, \tag{34}$$

where the expectation $\langle z_{tikd} \rangle = \varphi_{tikd}$.

3.3 Optimizing variational posterior $q(\boldsymbol{\Theta})$

Through the maximization of the ELBO with respect to $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\mu}', \boldsymbol{\lambda}'\}$, the variational posteriors of the VM distributions $q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ and $q\big(\boldsymbol{\mu}', \boldsymbol{\lambda}'\big)$ can be obtained by

$$\begin{aligned} & q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \\ & = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \mathrm{VM}\big(\boldsymbol{\mu}_{ikd} | \boldsymbol{m}_{ikd}^{*}, \beta_{ikd}^{*} \lambda_{ikd}\big) \mathcal{G}\big(\lambda_{ikd} | u_{ikd}^{*}, v_{ikd}^{*}\big), \end{aligned} \tag{35}$$

$$\begin{aligned} & q\big(\boldsymbol{\mu}', \boldsymbol{\lambda}'\big) \\ & = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \mathrm{VM}\big(\boldsymbol{\mu}'_{ikd} | \boldsymbol{m}'^{*}_{ikd}, \beta'^{*}_{ikd} \lambda'_{ikd}\big) \mathcal{G}\big(\lambda'_{ikd} | u'^{*}_{ikd}, v'^{*}_{ikd}\big), \end{aligned} \tag{36}$$

where the hyperparameters can be computed by

$$\beta_{ikd}^{*} = \| \beta_{ikd} \boldsymbol{m}_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k) \langle z_{tikd} \rangle \boldsymbol{x}_{td} \|, \tag{37}$$

$$\boldsymbol{m}_{ikd}^{*} = \frac{1}{\beta_{ikd}^{*}} \left(\beta_{ikd} \boldsymbol{m}_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k) \langle z_{tikd} \rangle \boldsymbol{x}_{td} \right), \tag{38}$$

$$u_{ikd}^{*} = u_{ikd} + \beta_{ikd}^{*} \bar{\lambda}_{ikd} \left(\frac{\partial}{\partial \beta_{ikd}^{*} \lambda_{ikd}} \ln I_0(\beta_{ikd}^{*} \bar{\lambda}_{ikd}) \right), \tag{39}$$

$$v^*_{ikd} = v_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k)\langle z_{tikd}\rangle \left(\frac{\partial}{\partial \lambda_{ikd}} \ln I_0\left(\bar{\lambda}_{ikd}\right) \right) + \beta_{ikd} \left(\frac{\partial}{\partial \beta_{ikd}\lambda_{ikd}} \ln I_0(\beta_{ikd}\bar{\lambda}_{ikd}) \right), \tag{40}$$

$$\beta'^*_{ikd} = \|\beta'_{ikd}\boldsymbol{m}'_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k)\langle 1 - z_{tikd}\rangle \boldsymbol{x}_{td}\|, \tag{41}$$

$$\boldsymbol{m}'^*_{ikd} = \frac{1}{\beta'^*_{ikd}} \left(\beta'_{ikd}\boldsymbol{m}'_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k)\langle 1 - z_{tikd}\rangle \boldsymbol{x}_{td} \right), \tag{42}$$

$$u'^*_{ikd} = u'_{ikd} + \beta'^*_{ikd}\bar{\lambda}'_{ikd} \left(\frac{\partial}{\partial \beta'^*_{ikd}\lambda'_{ikd}} \ln I_0(\beta'^*_{ikd}\bar{\lambda}'_{ikd}) \right), \tag{43}$$

$$v'^*_{ikd} = v'_{ikd} + \sum_{t=1}^{T} q(s_t = i, l_t = k)\langle 1 - z_{tikd}\rangle \left(\frac{\partial}{\partial \lambda'_{ikd}} \ln I_0(\bar{\lambda}'_{ikd}) \right) + \beta'_{ikd} \left(\frac{\partial}{\partial \beta'_{ikd}\lambda'_{ikd}} \ln I_0(\beta'_{ikd}\bar{\lambda}'_{ikd}) \right). \tag{44}$$

3.4 Optimizing variational posterior $q(S, L)$

Lastly, the joint variational posterior $q(S, L)$ is optimized (S represents the state indicator and L denotes the mixture component indicator) by maximizing the ELBO with respect to S and L

$$q(S, L) = \frac{1}{\Omega}\pi^*_{s_1} \prod_{t=1}^{T-1} a^*_{s_t s_{t+1}} \prod_{t=1}^{T} c^*_{s_t, l_t} p^*\left(\boldsymbol{y}_t | \Theta, z_t\right), \tag{45}$$

where

$$\pi^*_i = \exp\left\{ \Psi\left(\widehat{W}^{\pi}_i\right) - \Psi\left(\widehat{W}^{\pi}_i + \widetilde{W}^{\pi}_i\right) + \sum_{n=1}^{i-1} \left[\Psi\left(\widetilde{W}^{\pi}_n\right) - \Psi\left(\widehat{W}^{\pi}_n + \widetilde{W}^{\pi}_n\right) \right] \right\}, \tag{46}$$

$$a^*_{ij} = \exp\left\{ \Psi(\widehat{W}^A_{ij}) - \Psi(\widehat{W}^A_{ij} + \widetilde{W}^A_{ij}) + \sum_{n=1}^{j-1} \left[\Psi\left(\widetilde{W}^A_{in}\right) - \Psi\left(\widehat{W}^A_{in} + \widetilde{W}^A_{in}\right) \right] \right\}, \tag{47}$$

$$c^*_{ik} = \exp\left\{ \Psi\left(\widehat{W}^C_{ik}\right) - \Psi\left(\widehat{W}^C_{ik} + \widetilde{W}^C_{ik}\right) + \sum_{n=1}^{k-1} \left[\Psi\left(\widetilde{W}^C_{in}\right) - \Psi\left(\widehat{W}^C_{in} + \widetilde{W}^C_{in}\right) \right] \right\}, \tag{48}$$

$$\begin{aligned} p^*\left(\boldsymbol{y}_t | \Theta, z_t\right) = \exp\Bigg\{ & \sum_{d=1}^{D} \langle z_{tikd}\rangle \Big[\langle \lambda_{ikd}\boldsymbol{\mu}^{\mathrm{T}}_{ikd}\boldsymbol{x}_{td}\rangle \\ & - \ln 2\pi - \ln I_0\left(\bar{\lambda}_{ikd}\right) \\ & - \left(\frac{\partial}{\partial \lambda_{ikd}} \ln I_0\left(\bar{\lambda}_{ikd}\right) \right) \left(\langle \lambda_{ikd}\rangle - \bar{\lambda}_{ikd} \right) \Big] \\ & + \sum_{d=1}^{D} \langle 1 - z_{tikd}\rangle \Big[\langle \lambda'_{ikd}\boldsymbol{\mu}'^{\mathrm{T}}_{ikd}\boldsymbol{x}_{td}\rangle \\ & - \ln 2\pi - \left(\frac{\partial}{\partial \lambda'_{ikd}} \ln I_0(\bar{\lambda}'_{ikd}) \right) \left(\langle \lambda'_{jd}\rangle - \bar{\lambda}'_{ikd} \right) \\ & - \ln I_0\left(\bar{\lambda}'_{ikd}\right) \Big] \Bigg\}, \end{aligned} \tag{49}$$

where Ω in (45) is the normalizing constant and is given by

$$\Omega = q(X|\Phi^*) = \sum_{S,L} \pi^*_{s_1} \prod_{t=1}^{T-1} a^*_{s_t s_{t+1}} \prod_{t=1}^{T} c^*_{s_t, l_t} p^*\left(\boldsymbol{y}_t | \Theta, z_t\right). \tag{50}$$

It is noteworthy that (50) can be considered as the approximation to the likelihood of the model with optimized parameters Φ^*, as we compare (50) with (6).

Algorithm 1 provides the VB inference algorithm for learning the VM-NHMM-Fs model. This learning algorithm is guaranteed to converge as the ELBO in (17) is convex with respect to each variational posterior [4]. By monitoring the variation of the ELBO, we can easily discover the convergence status if the difference of the values of ELBO between two consecutive iterations is less than some predefined threshold.

Table 1 The parameters for generating the 3 relevant features for the 15-dimensional data set, where $S1$ and $S2$ indicate state 1 and state 2, respectively; n_k denotes the number of data points that are generated from the kth VM density, d represents the feature number

$S1$					$S2$				
k	n_k	d	$\boldsymbol{\mu}$	λ	k	n_k	d	$\boldsymbol{\mu}$	λ
1	750	1	(0.8575,0.5145)	3	1	750	1	(0.7071,0.7071)	8
		2	(0.3162,0.9487)	8			2	(0.8321,0.5547)	6
		3	(0.7649,0.6441)	16			3	(0.7682,0.6402)	10
2	750	1	(0.9751,0.2216)	10	2	750	1	(0.4472,0.8944)	20
		2	(0.5647,0.8253)	5			2	(0.6690,0.7433)	15
		3	(0.3511,0.9363)	20			3	(0.8480,0.5300)	6

Algorithm 1 The VB Inference of VM-NHMM-Fs.

1: Initialize hyperparameters: ϕ^{π}, ϕ^{A}, ϕ^{C}, u, v, u', v', β, β', $\boldsymbol{g}'$, and ζ.
2: Initialize variational distributions $q(s_1)$, $q(s_t, s_{t+1})$, $q(s_t, l_t)$ according to their prior distributions (13)$\sim$(14).
3: Calculate initial values of $\widehat{W}^{\pi}$, $\widehat{W}^{A}$, $\widehat{W}^{C}$, $\widetilde{W}^{\pi}$, $\widetilde{W}^{A}$ and $\widetilde{W}^{C}$ using Eqs. (24)$\sim$(29).
4: Calculate initial values of $\boldsymbol{\pi}$, A and C with Eqs. (46)$\sim$(48).
5: **repeat**
6: Optimize variational posterior distributions $q(\boldsymbol{\pi}')$, $q(A')$ and $q(C')$ using Eqs. (21)$\sim$(23).
7: Update the variational posterior of the feature saliency indicator $q(\boldsymbol{z})$ with Eq. (30).
8: Update the variational posteriors of the VM distributions $q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ and $q(\boldsymbol{\mu}', \boldsymbol{\lambda}')$ with Eqs. (35)$\sim$(36).
9: Update the variational posterior of indicator variables $q(S, L)$ with Eq. (45).
10: **until** Convergence
11: Compute the expected value of π_i' as $\langle \pi_i' \rangle = \widehat{W}_i^{\pi}/(\widehat{W}_i^{\pi} + \widetilde{W}_i^{\pi})$ and substitute it into (10) to obtain the estimated values of the initial state probabilities π_j.
12: Compute the expected value of c_{ik}' as $\langle c_{ik}' \rangle = \widehat{W}_{ik}^{C}/(\widehat{W}_{ik}^{C} + \widetilde{W}_{ik}^{C})$ and substitute it into (12) to obtain the estimated values of the mixing coefficients c_{ik}.
13: Detect the optimal number of hidden states N by eliminating the states with small probabilities (i.e. $\pi_i \to 0$).
14: Detect the optimal number of mixture components K in state i by eliminating the mixture components with small mixing coefficients (i.e. $c_{ik} \to 0$).

4 Experimental results

The proposed nonparametric HMM with localized feature selection (VM-NHMM-Fs) is evaluated through experiments on both synthetic and real-world time series or sequential data sets. We set the initial truncation values of N and K in our experiments as 20 and 30, respectively. The initial value of the hyperparameter ζ of the feature saliency is set to 0.5. The hyperparameters ϕ^{π}, ϕ^{A} and ϕ^{C} of the stick-breaking representation are all initialized to 0.5. The hyperparameters m and m' are initialized as the average of the data set. The other hyperparameters are initialized as: $(\beta, \beta', u, u', v, v') = (0.01, 0.01, 0.3, 0.3, 0.05, 0.05)$. We report the experimental results using the average performance of our model based on 20 runs for all experiments.

4.1 Experiments on synthetic sequential data

In this part, a synthetic sequential data set is generated to validate the effectiveness of the proposed learning approach to inferencing parameters and selecting important features for the proposed VM-NHMM-Fs.

Our synthetic sequential data set contains a sequence of 3000 data points that were generated based on 2 hidden states, where State 1 is used to generate the sequential observations at $t = 1 : 1500$, while state 2 is in charge of generating the sequential observations at $t = 1501 : 3000$. In each state, a mixture of two 3-dimensional VM densities corresponding to relevant features (i.e. we have 3 relevant features in total) was used as the emission density. The parameters that were adopted for generating the 3 relevant features are shown in Table 1. Then, we generated 12 irrelevant features according to a common VM distribution using parameters $\boldsymbol{\mu} = (0, 1)$ and $\lambda = 1$ and appended these features to the 3 relevant features to form a 15-dimensional data set.

To verify the "correctness" of the proposed VB learning algorithm, we compared the discrepancy between the true values of the parameters for generating the data set and the corresponding estimated values as in [3]. The comparison results of parameters for generating the synthetic sequential data set are demonstrated in Tables 2 and 3 , under state 1 and state 2, respectively. From these tables, we can see that the proposed VB inference algorithm can accurately estimate model parameters which illustrates the effectiveness of our VB algorithm.

Table 2 The comparison of the true and the estimated parameters by the proposed VM-NHMM-Fs under State 1 for the synthetic data set

$\min \boldsymbol{\mu}^{\top}\widehat{\boldsymbol{\mu}}$	$\text{avg}\boldsymbol{\mu}^{\top}\widehat{\boldsymbol{\mu}}$	$\max \frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lvert\lambda\rvert}$	$\text{avg}\frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lvert\lambda\rvert}$	$\max \frac{\lvert\pi-\widehat{\pi}\rvert}{\lvert\pi\rvert}$	$\text{avg}\frac{\lvert\pi-\widehat{\pi}\rvert}{\lvert\pi\rvert}$
0.997	0.998	0.002	0.004	0.001	0.003

Table 3 The comparison of the true and the estimated parameters by the proposed VM-NHMM-Fs under State 2 for the synthetic data set

$\min \boldsymbol{\mu}^{\top}\widehat{\boldsymbol{\mu}}$	$\text{avg}\boldsymbol{\mu}^{\top}\widehat{\boldsymbol{\mu}}$	$\max \frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lvert\lambda\rvert}$	$\text{avg}\frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lvert\lambda\rvert}$	$\max \frac{\lvert\pi-\widehat{\pi}\rvert}{\lvert\pi\rvert}$	$\text{avg}\frac{\lvert\pi-\widehat{\pi}\rvert}{\lvert\pi\rvert}$
0.998	0.997	0.003	0.002	0.001	0.002

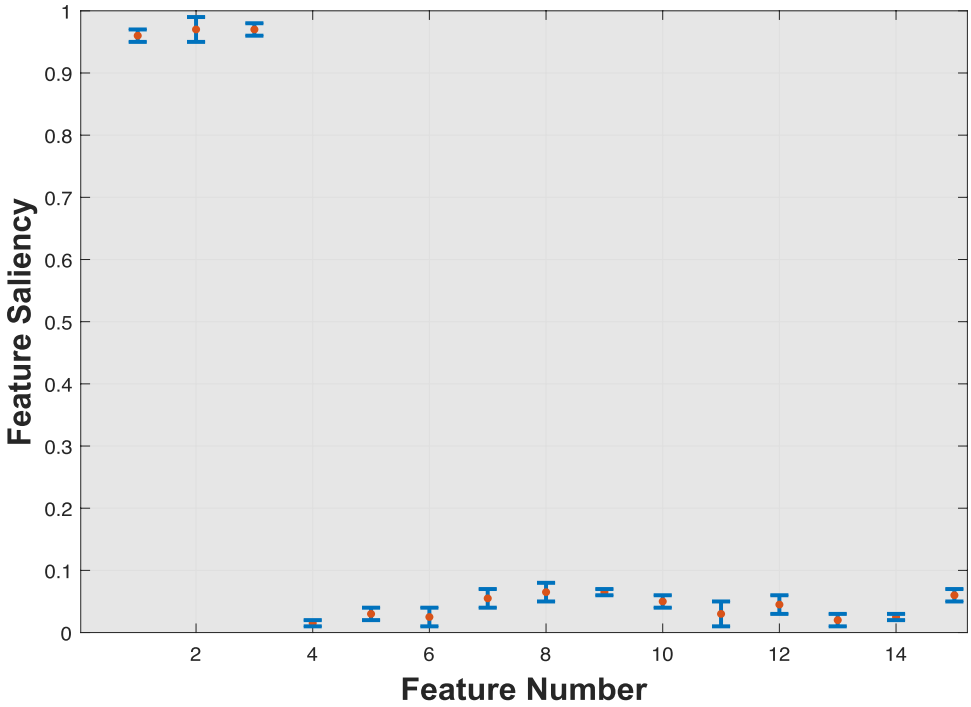

Fig. 2 Average feature saliences on the synthetic data set by VM-NHMM-FS plus and minus one standard deviation over 20 runs

Next, we test the performance of feature selection of our VM-NHMM-Fs on the synthetic data set. The results of feature selection in terms of feature saliency (i.e. the values of $\{\zeta_d\}$) are demonstrated in Fig. 2. According to the results shown in this figure, it is obvious that high degree of relevancies (i.e. above 0.9) have been assigned to the first three features while the remaining 12 features are considered as irrelevant features due to their low degrees of saliencies (i.e. close to 0). These results are consistent with the true settings of the synthetic sequential data set.

4.2 Experiments on real data sets

4.2.1 Data sets and experimental settings

In this part, the effectiveness of the proposed VM-NHMM-Fs was validated by conducting experiments on real sequential data sets in terms of unsupervised clustering applications. We adopted two real data sets from the UCI machine learning repository[1], including the gesture phase segmentation data set and the epileptic seizure recognition data set.

The gesture phase segmentation data set contains seven recorded videos consisted in a temporal segmentation of gestures (rest, preparation, stroke, hold and retraction) using Microsoft Kinect sensor. In our case, we teste the performance of VM-NHMM-Fs on three videos of this data set: A1 (1747 frames), A2 (1264 frames) and A3 (1834 frames), where each video includes the original version and a processed version. 50 features are extracted based on this data set, from which 18 features are obtained based on original videos and 32 features are extracted from processed videos.

The epileptic seizure recognition data set that we adopted is a pre-processed version of a data set regarding epileptic seizure detection as described in the UCI machine learning repository. It contains 11500 observations, where each observation consists of 178 data points, where each data point represents the value of the EEG observed at a different point in time. It contains five classes: (1) the EEG of seizure activity; (2) the EEG from the area where the tumor was located; (3) the EEG from the healthy brain area; (4) the EEG of the patient had their eyes closed; (5) the EEG of the patient had their eyes open.

In our experiment, these two data sets were L_2 normalized and then modeled by the proposed VM-NHMM-Fs. In order to demonstrate the advantages of our model, we compared it with other well-defined HMMs that employ different mixture models: the HMM with Gaussian mixture models (GMM-HMM) [21], the HMM with Gaussian mixture models and unsupervised feature selection (GMM-HMM-Fs) [43], the HMM with Dirichlet mixture model (DMM-HMM) [11], the HMM with inverted Dirichelt mixture model (IDMM-HMM) [30] and the HMM with VMF mixture models (VMF-HMM) [15]. Furthermore, to evaluate the importance of integrating feature selection in our model, we respectively applied the proposed model with localized feature selection (VM-NHMM-Fs) and without it (denoted by VM-NHMM). For the tested models, we adopted the same

[1] https://archive.ics.uci.edu.

Table 4 The average recognition performance over 20 runs by different approaches

Methods	Gesture phase	Epileptic seizure
GMM-HMM [21]	0.806 ± 0.007	0.716 ± 0.010
GMM-HMM-Fs [43]	0.821 ± 0.014	0.738 ± 0.018
DMM-HMM [11]	0.827 ± 0.009	0.732 ± 0.015
IDMM-HMM [30]	0.843 ± 0.012	0.749 ± 0.012
VMF-HMM [15]	0.861 ± 0.015	0.785 ± 0.013
VM-NHMM	0.875 ± 0.012	0.797 ± 0.009
VM-NHMM-Fs	0.903 ± 0.010	0.811 ± 0.012

parameter values as in their original papers. All tested models were implemented on the same data sets as described in our experiments.

In our experiment, we set the initial size of states for two data sets as $N = 20$, and the optimal number of states was automatically determined in the process of model learning. According to the results obtained by the proposed VM-NHMM-Fs, the gesture phase segmentation data set and the epileptic seizure recognition data set eventually converged to 3 and 2 states, respectively. For other tested approaches, the number of hidden states were set manually. Table 4 shows the recognition performance by different models on the two real data sets. As can be seen from this table, both VM-NHMM-Fs and VM-NHMM are able to outperform other HMM-based approaches with higher recognition accuracies for all data sets, which verified the merits of applying nonparametric VM-based HMMs for modeling gestures and EEG data. Another advantage of our approach is that, in contrast with other tested HMM-based approaches in which the number of clusters was determined through an extra evaluation step based on model selection scores, this number in our case was automatically determined during the inference procedure thanks to the nonparametric framework of Dirichlet process. According to Table 4, we may also notice the improvement of the performance when feature selection is integrated with VM-NHMM, by comparing the results of VM-NHMM-Fs with that of VM-NHMM.

The obtained feature saliencies of the 50-dimensional gesture phase data vectors of the resting phase by VM-NHMM-FS are shown in Fig. 3. It can be seen from this figure that there are 7 features that have obtained low degrees of relevance (i.e. saliencies are less than 0.5). Therefore, these features are considered as irrelevant features in the modeling process. On the other hand, the remaining features are considered as relevant features as they have high-level feature saliencies (i.e. greater than 0.5). Figure 4 illustrates the results of feature saliencies obtained by VM-NHMM-FS for the class of seizure activity of the epileptic seizure recognition data set. Based on this figure, different features have different contributions in the task of epileptic seizure recognition, where 22 of the 178 features have obtained relatively low saliencies (i.e. less than 0.5) and therefore have less contributions in data modeling.

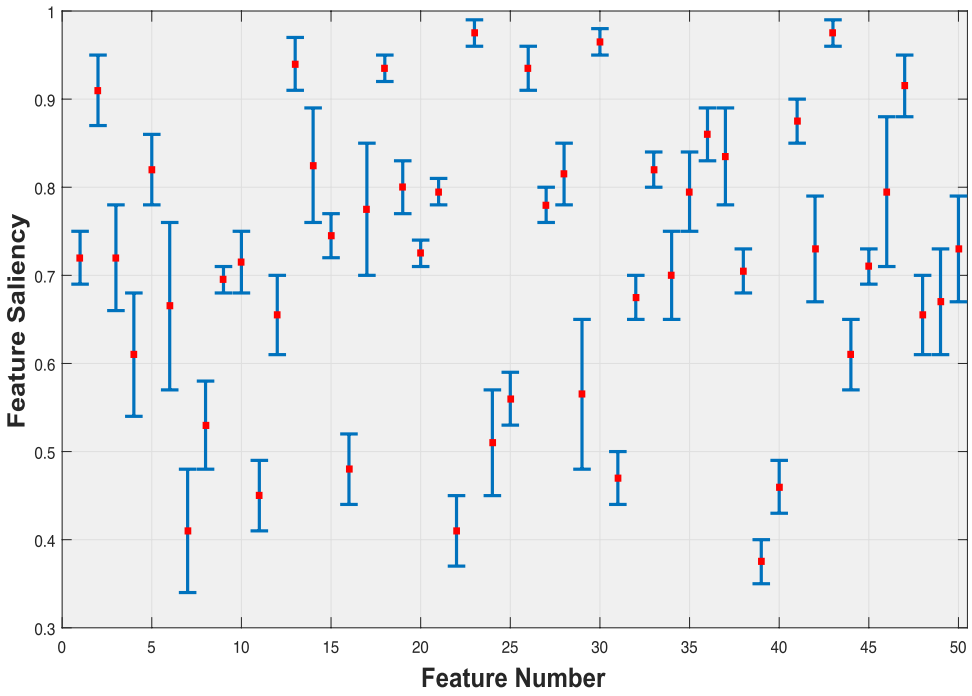

Fig. 3 Average feature saliences for the resting phase of the gesture phase segmentation data set by VM-NHMM-FS plus and minus one standard deviation over 20 runs

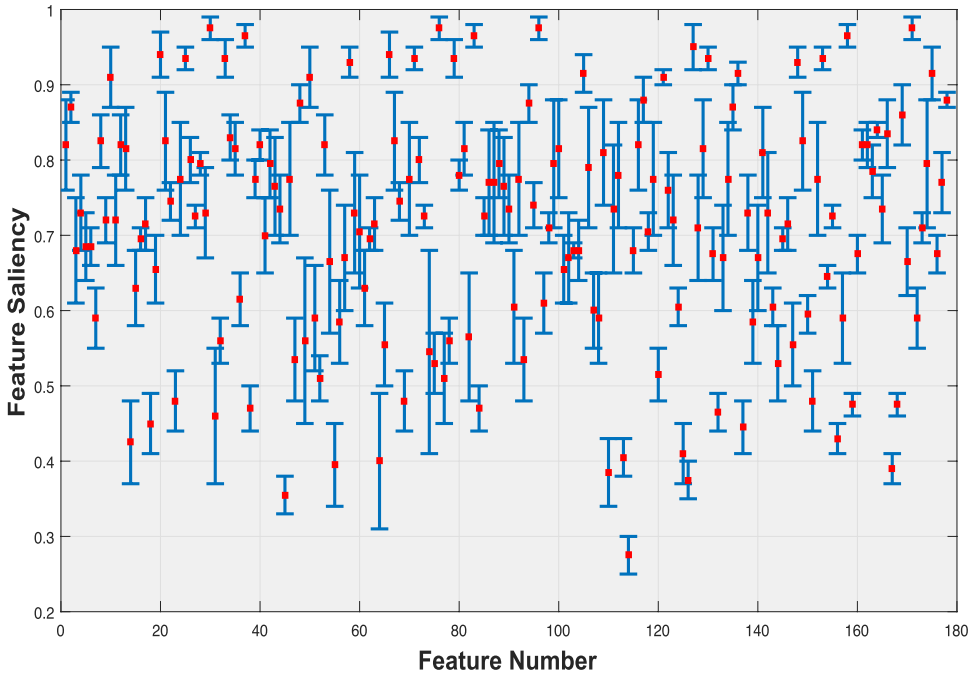

Fig. 4 Average feature saliences for the class of seizure activity of the epileptic seizure recognition data set by VM-NHMM-FS plus and minus one standard deviation over 20 runs

5 Conclusion

In this work, a nonparametric HMM has been proposed for modeling time series or sequential spherical data vectors. In our model, the emission distribution of each hidden state obeys a mixture of VM distributions which has shown better capability for modeling spherical data than other commonly used distributions (such as the Gaussian distribution). We constructed our NHMM by leveraging a Bayesian nonparametric DP framework, and therefore the amount of hidden states and the number of mixture components for each state can be automatically adjusted according to observed data set. In addition, to deal with high-dimensional data sets which may contain irrelevant or noisy features, an unsupervised localized feature selection method was incorporated with the proposed NHMM, which results in a unified framework that can simultaneously perform data modeling and feature selection. The proposed model was learned by developing an effective algorithm based on VB inference. The advantages of our model were demonstrated through both simulated and real-world data sets. Particularly, according to the experimental results, our model was able to outperform other tested HMM-based models by at least 4.2% in gesture recognition and at least 2.6% in epileptic seizure recognition.

One limitation of the proposed NHMM is that it is not very efficient for dealing with large-scale data sets. This is mainly caused by the batch learning strategy of the conventional VB inference adopted in our work. Thus, a possible future work is to extend the developed VB inference algorithm with stochastic variational Bayes (SVB) [18], which has shown its efficiency in learning over large data sets through stochastic optimization. Moreover, in recent years, deep learning techniques have been successfully applied in different fields owing to their promising capabilities of automatically extracting meaningful representations from observed data. Therefore, another interesting future work is to integrate deep neural networks (e.g. variational auto-encoder [23]) with the proposed NHMM to improve its performance by leveraging the more representative features learned by these deep learning techniques.

Acknowledgements The completion of this work was supported by the National Natural Science Foundation of China (61876068).

Data availability statement The data sets analysed during the current study are available in the UCI Machine Learning Repository https://archive.ics.uci.edu.

References

1. Asilian Bidgoli A, Ebrahimpour-komleh H, Rahnamayan S (2021) A novel binary many-objective feature selection algorithm for multi-label data classification. Int J Mach Learn Cybern 12:2041–2057
2. Aytekin C, Ni X, Cricri F, Aksu E (2018) Clustering and unsupervised anomaly detection with l_2 normalized deep auto-encoder representations. In: 2018 international joint conference on neural networks (IJCNN), pp 1–6
3. Banerjee A, Dhillon I, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res 6:1345–1382
4. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

5. Blei DM, Jordan MI (2005) Variational inference for Dirichlet process mixtures. Bayesian Anal 1:121–144
6. Blei DM, Kucukelbir A, Mcauliffe J (2017) Variational inference: a review for statisticians. J Am Stat Assoc 112(518):859–877
7. Calderara S, Prati A, Cucchiara R (2011) Mixtures of von Mises distributions for people trajectory shape analysis. IEEE Trans Circ Syst Video Technol 21(4):457–471
8. Chatzis SP, Kosmopoulos DI (2011) A variational Bayesian methodology for hidden Markov models utilizing student's-t mixtures. Pattern Recogn 44(2):295–306
9. Ding N, Ou Z (2010) Variational nonparametric Bayesian hidden markov model. In: 2010 IEEE international conference on acoustics, speech and signal processing, pp 2098–2101
10. Dokeroglu T, Deniz A, Kiziloz HE (2021) A robust multiobjective harris' hawks optimization algorithm for the binary classification problem. Knowl-Based Syst 227(107):219
11. Epaillard E, Bouguila N (2019) Variational Bayesian learning of generalized Dirichlet-based hidden Markov models applied to unusual events detection. IEEE Trans Neural Netw 30(4):1034–1047
12. Fan W, Bouguila N (2020) Spherical data clustering and feature selection through nonparametric Bayesian mixture models with von Mises distributions. Eng Appl Artif Intell 94(103):781
13. Fan W, Bouguila N, Ziou D (2011) Unsupervised anomaly intrusion detection via localized Bayesian feature selection. In: 2011 IEEE 11th international conference on data mining (ICDM), pp 1032–1037
14. Fan W, Bouguila N, Du J, Liu X (2019) Axially symmetric data clustering through Dirichlet process mixture models of Watson distributions. IEEE Trans Neural Netw Learn Syst 30(6):1683–1694
15. Fan W, Yang L, Bouguila N, Chen Y (2020) Sequentially spherical data modeling with hidden Markov models and its application to fMRI data analysis. Knowl-Based Syst 206(106):341
16. Fan W, Yang L, Bouguila N (2021) Unsupervised grouped axial data modeling via hierarchical Bayesian nonparametric models with Watson distributions. IEEE Trans Pattern Anal Mach Intell 2021:1–1. https://doi.org/10.1109/TPAMI.2021.3128271
17. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
18. Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. J Mach Learn Res 14(1):1303–1347
19. Illingworth CJR, Roy S, Beale MA, Tutill HJ, Williams R, Breuer J (2017) On the effective depth of viral sequence data. Virus Evol 3:2
20. Javidi MM (2021) Feature selection schema based on game theory and biology migration algorithm for regression problems. Int J Mach Learn Cybern 12:303–342
21. Ji S, Krishnapuram B, Carin L (2006) Variational Bayes for continuous hidden Markov models and its application to active learning. IEEE Trans Pattern Anal Mach Intell 28(4):522–532
22. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Mach Learn 37(2):183–233
23. Kingma DP, Welling M (2014) Auto-encoding variational Bayes. In: ICLR
24. Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26(9):1154–1166
25. Ley C, Verdebout T (2018) Applied directional statistics: modern methods and case studies. Chapman and Hall/CRC, Hoboken
26. Li J, Cheng K, Wang S, Morstatter F, Trevino R, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Comput Surv 50(6):94
27. Li Y, Dong M, Hua J (2009) Simultaneous localized feature selection and model detection for Gaussian mixtures. IEEE Trans Pattern Anal Mach Intell 31(5):953–960
28. Mabrouk AB, Zagrouba E (2018) Abnormal behavior recognition for intelligent video surveillance systems. Expert Syst Appl 91:480–491
29. Mardia KV, Jupp PE (2000) Directional statistics. Wiley, USA
30. Nasfi R, Amayri M, Bouguila N (2020) A novel approach for modeling positive vectors with inverted Dirichlet-based hidden Markov models. Knowl Based Syst 192(105):335
31. Pigou L, Den Oord AV, Dieleman S, Van Herreweghe M, Dambre J (2018) Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. Int J Comput Vis 126:430–439
32. Qiu Z, Shen H (2017) User clustering in a dynamic social network topic model for short text streams. Inf Sci 414:102–116
33. Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
34. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):267–296
35. Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sin 4:639–650
36. Sra S, Karp D (2013) The multivariate Watson distribution: Maximum-likelihood estimation and other aspects. J Multivar Anal 114:256–269
37. Taghia J, Leijon A (2016) Variational inference for Watson mixture model. IEEE Trans Pattern Anal Mach Intell 38(9):1886–1900
38. Taghia J, Ma Z, Leijon A (2014) Bayesian estimation of the von Mises-fisher mixture model with variational inference. IEEE Trans Pattern Anal Mach Intell 36(9):1701–1715
39. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. J Am Stat Assoc 101(476):1566–1581
40. Tubishat M, Ja'afar S, Alswaitti M, Mirjalili S, Idris N, Ismail MA, Omar MS (2021) Dynamic salp swarm algorithm for feature selection. Expert Syst Appl 164(113):873
41. Volant S, Berard C, Martinmagniette M, Robin S (2014) Hidden markov models with mixtures as emission distributions. Stat Comput 24(4):493–504
42. Zheng Y, Jeon B, Sun L, Zhang J, Zhang H (2018) Student's t-hidden Markov model for unsupervised learning using localized feature selection. IEEE Trans Circuits Syst Video Technol 28(10):2586–2598
43. Zhu H, He Z, Leung H (2012) Simultaneous feature and model selection for continuous hidden markov models. IEEE Signal Process Lett 19(5):279–282

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.