Received: 8 July 2019 | Revised: 1 July 2020 | Accepted: 13 July 2020

DOI: 10.1049/ipr2.12154

IET Image Processing

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Online variational inference on finite multivariate Beta mixture models for medical applications

Narges Manouchehri | **Meeta Kalra** | **Nizar Bouguila**

Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada

Correspondence
Narges Manouchehri, Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada - H3G 1M8.
Email: narges.manouchehri@mail.concordia.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN/6656-2017

Abstract

Technological advances led to the generation of large scale complex data. Thus, extraction and retrieval of information to automatically discover latent pattern have been largely studied in the various domains of science and technology. Consequently, machine learning experienced tremendous development and various statistical approaches have been suggested. In particular, data clustering has received a lot of attention. Finite mixture models have been revealed to be one of the flexible and popular approaches in data clustering. Considering mixture models, three crucial aspects should be addressed. The first issue is choosing a distribution which is flexible enough to fit the data. In this paper, a model based on multivariate Beta distributions is proposed. The two other challenges in mixture models are estimation of model's parameters and model complexity. To tackle these challenges, variational inference techniques demonstrated considerable robustness. In this paper, two methods are studied, namely, batch and online variational inferences and the models are evaluated on four medical applications including image segmentation of colorectal cancer, multi-class colon tissue analysis, digital imaging in skin lesion diagnosis and computer aid detection of Malaria.

1 | INTRODUCTION

Over the past decades, fast progress of computational power and data storage yield a great deal of complex data and machine learning methods experienced considerable development to recognize critical information from data efficiently and automatically with minimal human interaction. In order to cover the wide variety of data such as text, image and video and problem types exhibited across different domains, a diverse array of machine learning algorithms have been developed [1]. Many algorithms focus on image processing and computer vision as techniques of electronics engineering.

A critical scientific and practical goal to the majority of the algorithms is to characterize their capabilities and robustness. Supervised learning systems have been widely used over the past years. Deep learning platforms [2, 3] have been demonstrated to outperform previous supervised machine learning techniques in several fields. Convolutional neural networks [4] and deep belief networks [5] are some examples of currently remarkable techniques in some applications such as image analysis [6], emotion detection [7], object detection [8–11], synthetic aperture radar image analysis [12, 13], remote sensing [14], Internet of Things [15], smart cities [16]. Similarly, modern medical imaging have witnessed admirable progresses and became one of the attention-grabbing domains in research and technology. Consequently, statistical modelling has been applied successfully in this domain and achieved state-of-the-art performance in image segmentation and computer-aided detection (CAD) to assist professionals in the interpretation of medical images, digital pathology and other medical datasets [17]. Due to the increasing digitization in medical image results [18] and prompt progression in artificial intelligence (AI) and machine learning (ML), various methods have been proposed [18]. However, the nature of medical data and some needs of healthcare team in making decision led to a limited success in applying the current algorithms in routine clinical cases, [19]. It should be noted that with some deep learning platforms we can achieve good results in classification tasks in various medical domains such as brain image analysis [20], pathological image analysis [21], cardiac image analysis [22], breast histology images

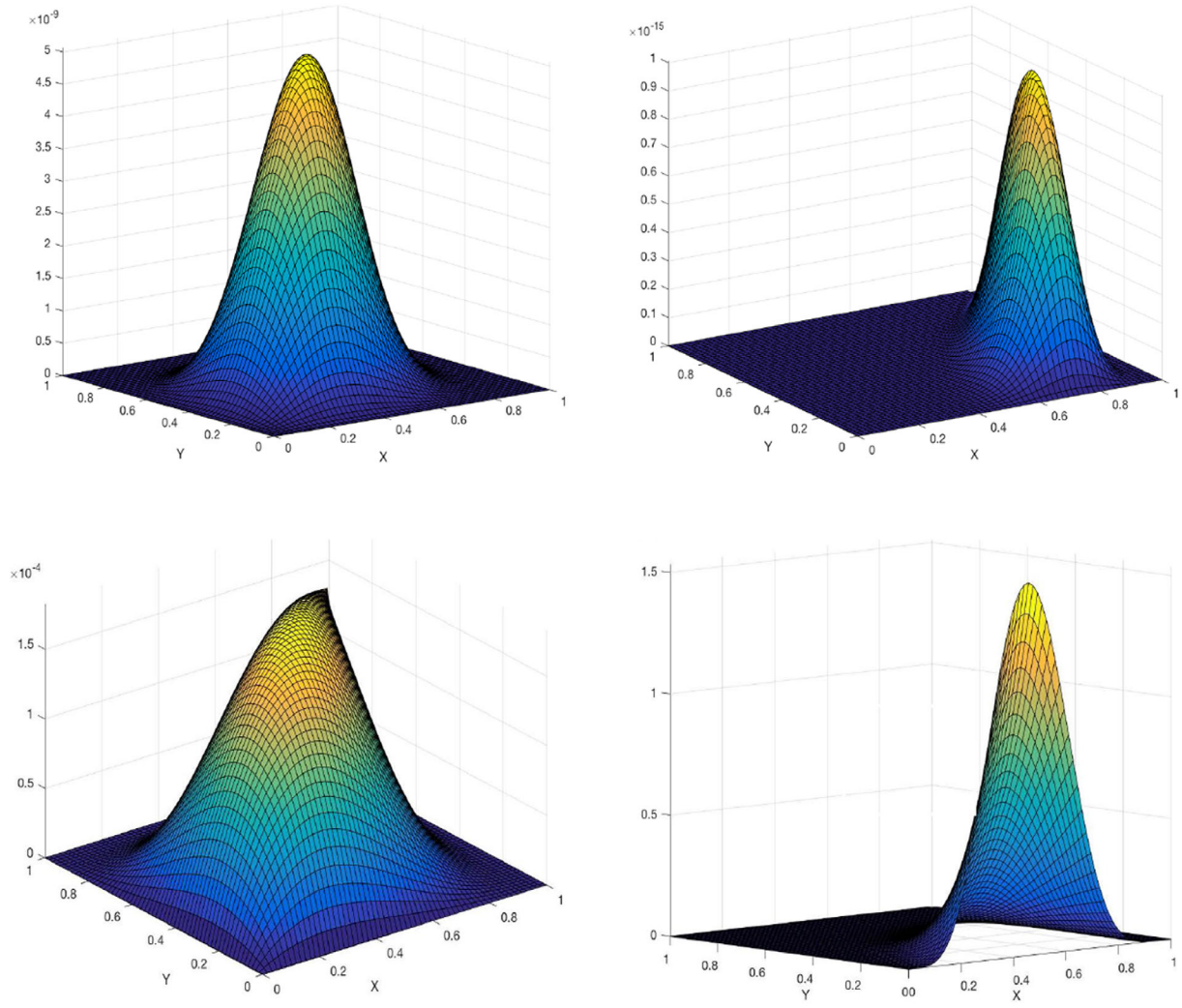

FIGURE 1 Four examples of multivariate Beta distributions

analysis [23], blood cell analysis [24], liver tumour analysis [25]. However, they may cause some failure as they are unpredictable and unexplainable [26–28]. It should be emphasized that deep learning models need large scale labelled data for training and the publicly available datasets are limited as confidentiality is a principle rule in healthcare. However, this is not the only issue, but medical data labelling is a great obstacle as it could be performed just by professional physicians and need sophisticated amount of budget, time and skill. It is noteworthy that the nature of medical data is heterogeneous and to arrive at a better decision, the model should have the potential and ability to deal with various types of data such as patient history, images, videos and signals, simultaneously.

These characteristics and demand, motivated us to focus on unsupervised models of machine learning as label-free approaches. Clustering methods especially finite mixture models are one of the best known methods to model heterogeneous data which includes multiple distributions [29]. The first challenging aspect which should be carefully addressed is choosing the most proper distributions that best represent the corresponding components of mixture accurately when modelling data. Gaussian mixture models (GMM) have been widely adopted in various applications [30]. However, in recent works other alternatives such as Dirichlet [31, 32], generalized Dirichlet [33–35] demonstrated considerable flexibility and high potential to describe non-Gaussian data. Hence, in our paper we focus on multivariate Beta mixture models which are developed based on a very flexible distribution which does not have a constant shape and is appropriate to be used to model data skewness. Furthermore, considering its bounded nature, it fits better compactly supported data. Figure 1 illustrates the high potential of this distribution.

To design a clustering algorithm, the parameters estimation is a crucial step and has a significant impact on the performance of model learning. The majority of parameter estimation methods apply either deterministic or Bayesian techniques. The former one is based on classic maximum likelihood (ML) inference and optimizing the model likelihood function via expectation–maximization (EM) [36] framework. However,

this method is sensitive to initialization and carry disadvantages such as over-fitting. To avoid such drawbacks, Bayesian techniques have been proposed. In this improved method, a prior knowledge is applied in a principled way and the parameter uncertainty is then marginalized by Laplace's approximation or Markov chain Monte Carlo (MCMC) simulation techniques [37, 38]. Unfortunately, we face some issues in Bayesian inference. For instance, Laplace's approximation is generally imprecise and MCMC techniques are computationally expensive. Recently, several research efforts focused on variational inference [39] as a preferable and efficient alternative technique for the learning of statistical models. Indeed, it can be expressed as an effective compromise between deterministic and Bayesian approaches. Variational inference is based on approximating the model posterior distribution which is achieved by minimizing the Kullback–Leibler (Kl) divergence between the true posterior and an approximating distribution. Another crucial issue when using mixture models is defining model structure or the best number of mixture components that describes the data perfectly without over-fitting or under-fitting. Some model selection techniques such as MML or MDL [31, 40, 41] have been considered. However, they are time-consuming since they have to evaluate a given selection criterion for several numbers of mixture components and such high computational cost limited their applications. One of the advantages of variational inference is that it automatically determines the number of mixture components as part of the Bayesian inference procedure [42, 43]. Variational learning can be performed online [44] which is mainly motivated by the fact that such algorithm allows data instances to be processed in a sequential way, which is important for large-scale data and real-time applications. This technique is significantly faster than traditional variational learning. In this paper, we propose two novel algorithms for batch and online variational learning based on multivariate Beta mixture models. We evaluate the performance of our proposed frameworks by exploring challenging medical applications and the results are compared with batch and online variational learning for Gaussian mixture models.

The structure of the rest of this paper is as follows; Section 2 is devoted to the description of finite multivariate Beta mixture model. Sections 3 and 4 describe the batch and online variational learning algorithms, respectively. We present the experimental results in Section 5 considering four real-world applications. Finally, we conclude in Section 6.

2 | FINITE MULTIVARIATE BETA MIXTURE MODEL

In this section, we give a brief description of finite multivariate Beta mixture models. Let us assume that an observation following a multivariate Beta (MB) distribution [45, 46] is defined by $\vec{X}_i = (x_{i1}, \dots, x_{iD})$ as a D-dimensional vector where all its elements are positive and less than one. $\Gamma(.)$ denotes the Gamma function. The probability density function of MB is expressed by (1).

$\vec{\alpha}_j = (\alpha_{j1}, \dots, \alpha_{jD})$ is shape parameter such that $\alpha_{jl} > 0$ for $l = 1, \dots, D$ and $\mid \alpha_j \mid = \sum_{l=1}^{D} \alpha_{jl}$.

$$p(\vec{X}_i \mid \vec{\alpha}_j) = c \frac{\prod_{l=1}^{D} x_{il}^{\alpha_{jl}-1}}{\prod_{l=1}^{D} (1-x_{il})^{(\alpha_{jl}+1)}} \left[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1-x_{il})}\right]^{-|\alpha_j|} \tag{1}$$

$$c = \frac{\Gamma(\alpha_{j1} + \cdots + \alpha_{jD})}{\Gamma(\alpha_1) \dots \Gamma(\alpha_{jD})} = \frac{\Gamma(\mid \alpha_j \mid)}{\prod_{l=1}^{D} \Gamma(\alpha_{jl})}$$

Let us consider a set of N independent identically distributed vectors $\mathcal{X} = \{\vec{X}_1, \dots, \vec{X}_N\}$ which are generated from multivariate Beta mixture models and composed of M different clusters. Thus, multivariate Beta mixture model is represented by:

$$p\left(\vec{X}_i \mid \vec{\pi}, \vec{\alpha}\right) = \sum_{j=1}^{M} \pi_j p\left(\vec{X}_i \mid \vec{\alpha}_j\right) \tag{2}$$

where $\vec{\pi} = (\pi_1, \dots, \pi_M)$ is the set of mixing coefficients with two constraints $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j >= 0$. $\vec{\alpha}_j$ and π_j are shape parameter and weight of component j where $j = 1, \dots, M$. So, the likelihood function for N samples is,

$$p\left(\mathcal{X} \mid \vec{\pi}, \vec{\alpha}\right) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j p\left(\vec{X}_i \mid \vec{\alpha}_j\right)\right] \tag{3}$$

Four examples of multivariate Beta mixture models (MBMM) are shown in Figure 2.

In mixture models, we define an auxiliary variable $\mathcal{Z}$ to allocate each sample to one of the M components. Thus, we introduce $\vec{Z}_i = (Z_{i1}, \dots, Z_{iM})$ where Z_{ij} is a binary random variable such that $Z_{ij} = 1$ if $\vec{X}_i$ belongs to the specific cluster j and 0, otherwise. The distribution of $\mathcal{Z} = \{\vec{Z}_1, \dots, \vec{Z}_N\}$ as a set of "membership vectors" is specified by (4) in terms of the mixing coefficients $\vec{\pi}$ [47].

$$p\left(\mathcal{Z} \mid \vec{\pi}\right) = \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{4}$$

Thus, the conditional probability of the data given $\mathcal{Z}$ is,

$$p\left(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}\right) = \prod_{i=1}^{N} \prod_{j=1}^{M} p\left(\vec{X}_i \mid \vec{\alpha}_j\right)^{Z_{ij}} \tag{5}$$

3 | BATCH VARIATIONAL LEARNING

Variational approaches have been widely applied previously to approximate posterior distributions of a variety of statistical models. In this section as the first step, we develop

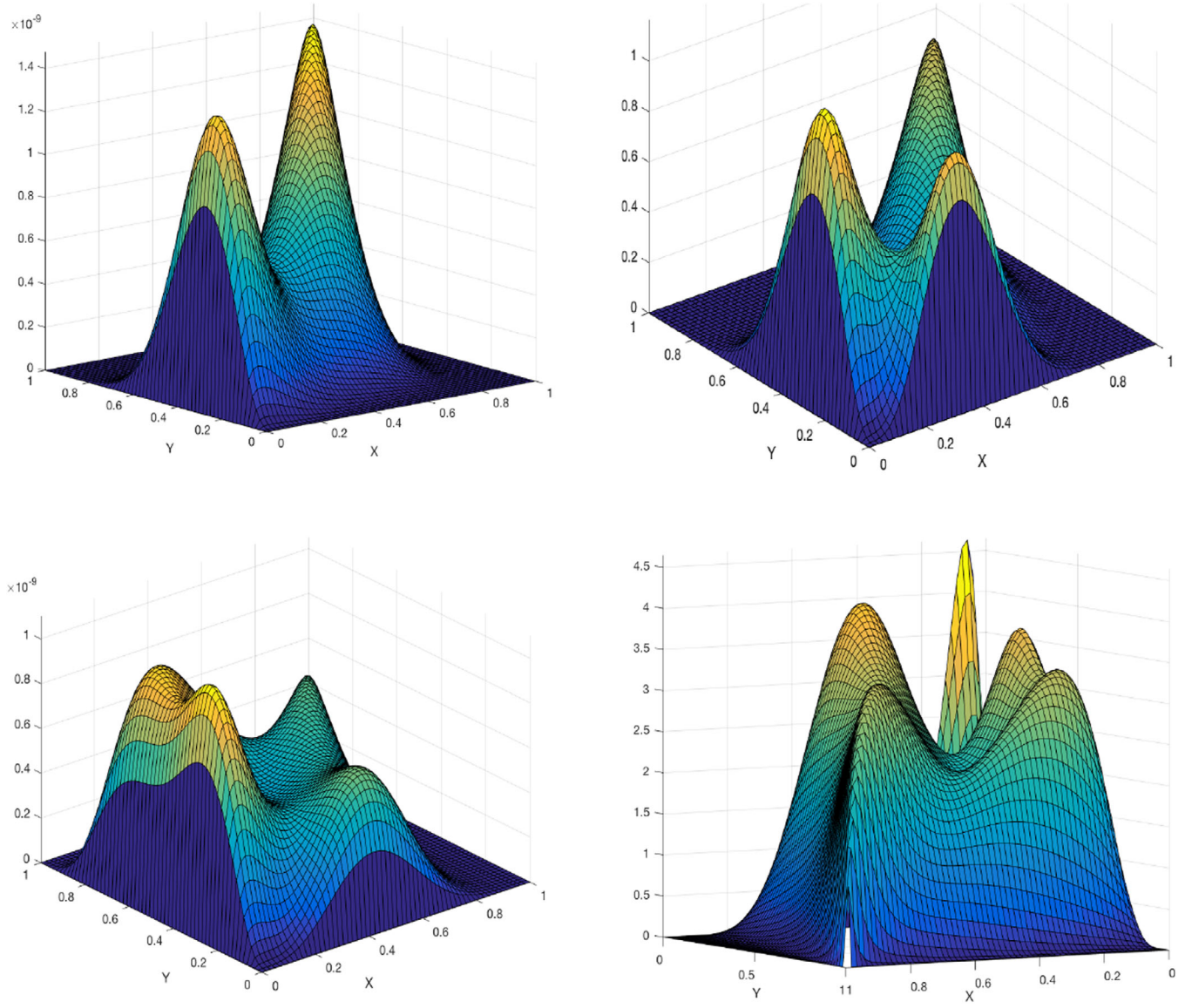

FIGURE 2 Four examples of MBMM with different components

a batch variational inference framework for learning finite MBMM. Our main objective is to develop an optimized method which is capable enough to estimate the parameters of mixture model and determine its structure and complexity simultaneously.

3.1 | Prior specification

A crucial challenge in the case of variational learning is placing prior distributions over parameters. To simplify this approach, we consider a conjugate prior for the $\vec{\alpha}$ parameters. Unfortunately, a conjugate prior does not exist. In this case, we adopt a Gamma prior as an approximation assuming that the parameters are statistically independent [48, 49]. So, the probability density function of α_{jl} is described by (6). u_{jl} and ν_{jl} are positive hyperparameters.

$$p(\alpha_{jl}) = \mathcal{G}(\alpha_{jl} \mid u_{jl}, \nu_{jl}) = \frac{\nu_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{6}$$

The model parameters $\vec{\alpha}$ are given by:

$$p(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} p(\alpha_{jl}) \tag{7}$$

Thus, the joint distribution of all random variables is given by:

$$p(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}) = p(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}) p(\mathcal{Z} \mid \vec{\pi}) p(\vec{\alpha}) \tag{8}$$

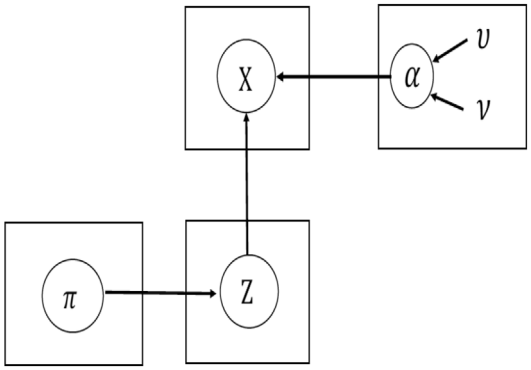

FIGURE 3 Graphical model of the finite multivariate Beta mixture

$$= \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\frac{\Gamma(\mid \alpha_j \mid)}{\prod_{l=1}^{D}\Gamma(\alpha_{jl})} \times \frac{\prod_{l=1}^{D} x_{il}^{\alpha_{jl}-1}}{\prod_{l=1}^{D}(1-x_{il})^{(\alpha_{jl}+1)}}\right.$$

$$\left. \times \left[1 + \sum_{l=1}^{D}\frac{x_{il}}{(1-x_{il})}\right]^{-|\alpha_j|}\right]^{Z_{ij}}$$

$$\times \prod_{i=1}^{N}\left[\prod_{j=1}^{M}\pi_j^{Z_{ij}}\right] \times \prod_{j=1}^{M}\prod_{l=1}^{D}\frac{\nu_{jl}^{u_{jl}}}{\Gamma(u_{jl})}\alpha_{jl}^{u_{jl}-1}e^{-\nu_{jl}\alpha_{jl}}$$

A graphical representation of this model is shown in Figure 3. Symbols in circles denote random variables; otherwise, they denote model parameters. The conditional dependencies of the variables are represented by the arcs.

3.2 | Learning algorithm

In order to estimate the parameters of model and select a correct number of components, we estimate the mixing coefficient $\vec{\pi}$ by maximizing the marginal likelihood $p(\mathcal{X} \mid \vec{\pi})$ expressed by (9).

$$p(\mathcal{X} \mid \vec{\pi}) = \sum_{\mathcal{Z}}\int p(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi})\mathrm{d}\vec{\alpha} \tag{9}$$

As the marginalization of this equation is intractable, we apply variational inference [50] to calculate the lower bound on $p(\mathcal{X} \mid \vec{\pi})$. The variational lower bound $\mathcal{L}$ of the logarithm of the marginal likelihood $p(\mathcal{X} \mid \vec{\pi})$ is defined by:

$$\mathcal{L}(\mathcal{Q}) = \int \mathcal{Q}(\Theta)\ln\left(\frac{p(\mathcal{X}, \Theta \mid \vec{\pi})}{\mathcal{Q}(\Theta)}\right)\mathrm{d}\Theta \tag{10}$$

where $\Theta = \{\mathcal{Z}, \vec{\alpha}, \vec{\pi}\}$ and $\mathcal{Q}(\Theta)$ is an approximation to the true posterior distribution $p(\Theta \mid \mathcal{X}, \vec{\pi})$. This approximation is determined by computation of KL divergence between $\mathcal{Q}(\Theta)$ and $p(\Theta \mid \mathcal{X}, \vec{\pi})$ defined by (11).

$$\mathrm{KL}(\mathcal{Q} \mid\mid P) = -\int \mathcal{Q}(\Theta)\ln\left(\frac{p(\Theta \mid \mathcal{X}, \vec{\pi})}{\mathcal{Q}(\Theta)}\right)\mathrm{d}\Theta \tag{11}$$

$$\mathrm{KL}(\mathcal{Q} \mid\mid P) = \ln\, p(\mathcal{X} \mid \vec{\pi}) - \mathcal{L}(\mathcal{Q}) \tag{12}$$

The KL divergence is the representation of the dissimilarity between the true posterior and its approximation. As $\mathrm{KL}(\mathcal{Q} \mid\mid P) \geq 0$, the $\mathrm{KL}(\mathcal{Q} \mid\mid P)$ is zero when $\mathcal{Q}(\Theta) = p(\Theta \mid \mathcal{X})$. Considering above mentioned equations, it is obvious that $\mathcal{L}(\mathcal{Q}) \leq \ln\, p(\mathcal{X} \mid \vec{\pi})$, thus $\mathcal{L}(\mathcal{Q})$ is a lower bound on $\ln\, p(\mathcal{X} \mid \vec{\pi})$. So, by maximizing the lower bound, the KL divergence is minimized and hence the true posterior distribution is approximated. Consequently, we consider a restricted and tractable family of distributions $\mathcal{Q}(\Theta)$ which are flexible enough to properly approximate the true posterior distribution. We apply common method, namely, mean field theory to adopt factorization assumptions for restricting the form of $\mathcal{Q}(\Theta)$. Subsequently, the posterior distribution $\mathcal{Q}(\Theta)$ can be factorized [48] such that,

$$\mathcal{Q}(\Theta) = \mathcal{Q}(\mathcal{Z})\mathcal{Q}(\vec{\alpha})\mathcal{Q}(\vec{\pi}) \tag{13}$$

We find the variational solution for $\mathcal{L}(\mathcal{Q})$ with respect to each of the parameters to maximize the lower bound and for a specific parameter s, the optimal solution can be expressed by:

$$\ln \mathcal{Q}_s^*(\Theta_s) = \langle \ln p(\mathcal{X}, \Theta)\rangle_{i\neq s} \tag{14}$$

By taking the exponential from both sides of this equation and normalizing, we can get:

$$\mathcal{Q}_s(\Theta_s) = \frac{\exp\left\langle \ln\, p(\mathcal{X}, \Theta)\right\rangle_{i\neq s}}{\int \exp\left\langle \ln\, p(\mathcal{X}, \Theta)\right\rangle_{i\neq s}\mathrm{d}\Theta} \tag{15}$$

$< \cdot >_{i\neq s}$ is the expectation with respect to all the parameters other than Θ_s.

The solutions for the optimal variational posteriors as derived in Appendix A are given by:

$$\mathcal{Q}(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{16}$$

$$\mathcal{Q}(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D}\mathcal{G}\left(\alpha_{jl} \mid u_{jl}^*, \nu_{jl}^*\right) \tag{17}$$

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M}\tilde{r}_{ij}} \tag{18}$$

$$\tilde{r}_{ij} = \exp\left\{\ln \pi_j + \tilde{R}_j + \sum_{l=1}^{D}\left(\overline{\alpha}_{jl} - 1\right)\ln x_{il}\right. \tag{19}$$

$$\left. -\sum_{l=1}^{D}\left(\overline{\alpha}_{jl} + 1\right)\ln(1-x_{il}) - \mid \overline{\alpha}_j \mid \ln\left[1 + \sum_{l=1}^{D}\frac{x_{il}}{(1-x_{il})}\right]\right\}$$

where $\tilde{R}_j$ is as follows based on [51] and its calculation is presented in Appendix A.

$$u_{jl}^{*} = u_{jl} + \varphi_{jl}, \quad v_{jl}^{*} = v_{jl} - \vartheta_{jl} \tag{20}$$

$$\varphi_{jl} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \overline{\alpha}_{jl} \left[\psi\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right), -\psi\left(\overline{\alpha}_{jl}\right), + \sum_{s \neq l}^{D} \psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right) \times \overline{\alpha}_{js} \left(\left\langle \ln \alpha_{js} \right\rangle - \ln \overline{\alpha}_{js} \right) \right] \tag{21}$$

$$\vartheta_{jl} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\ln x_{il} - \ln(1 - x_{il}) - \ln \left[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})} \right] \right] \tag{22}$$

$\psi(.)$ and $\psi'(.)$ in the above equations represent the digamma and trigamma functions. The expectation of values mentioned in the equations above is given by,

$$\langle Z_{ij} \rangle = r_{ij} \tag{23}$$

$$\overline{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}^{*}}{v_{jl}^{*}} \tag{24}$$

$$\left\langle \ln \alpha_{jl} \right\rangle = \psi\left(u_{jl}^{*}\right) - \ln v_{jl}^{*} \tag{25}$$

$$\left\langle \left(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl} \right)^2 \right\rangle = \left[\psi\left(u_{jl}^{*}\right) - \ln u_{jl}^{*} \right]^2 + \psi'\left(u_{jl}^{*}\right) \tag{26}$$

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{27}$$

In variational learning, we trace the convergence systematically by monitoring the variational lower bound during the re-estimation step. Indeed, at each step of the iterative updating procedure, the value of $\mathcal{L}(Q)$ should not rise. Thus, we terminate optimization when the lower bound increases more than a threshold compared to previous estimated value. The lower bound in (10) is evaluated as follows which is explained in details in Appendix A :

$$\mathcal{L}(Q) = \sum_{\mathcal{Z}} \int Q(\mathcal{Z}, \vec{\alpha}) \ln \left(\frac{p(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi})}{Q(\mathcal{Z}, \vec{\alpha})} \right) \mathrm{d}\vec{\alpha} \tag{28}$$

The complete algorithm of batch variational learning can be summarized in Algorithm 1.

ALGORITHM 1 Batch variational framework for MBMM

1. Choose a large initial number of components M.
2. Initialize values for u_{jd} and v_{jd}.
3. Initialize the value of r_{ij} using K-Means algorithm.
4. **repeat**
5. The variational E-step:
6. Estimate the expected values by equations (23) to (27).
7. The variational M-step:
8. Update $Q(\mathcal{Z})$ and $Q(\vec{\alpha})$ by estimating r_{tj} from (16) and (17).
9. **until** Convergence criterion is reached.

4 | ONLINE VARIATIONAL LEARNING OF MULTIVARIATE BETA MIXTURE MODELS

In this subsection, we extend the classic variational inference approach [49] to online settings for learning multivariate Beta mixture model by adopting the framework proposed in [44] as in real-world, observations arrive in an online manner. Thus, we assume that a specific amount of data are observed defined by t, such that their corresponding lower bound is defined by [52]:

$$\mathcal{L}^{(t)}(Q) = \frac{N}{t} \sum_{i=1}^{t} \int Q(\alpha) \mathrm{d}\alpha \sum_{\vec{Z}_i} Q(\vec{Z}_i) \ln \left[\frac{p(\vec{X}_i, \vec{Z}_i \mid \alpha)}{Q(\vec{Z}_i)} \right] + \int Q(\alpha) \ln \left[\frac{p(\alpha)}{Q(\alpha)} \right] \mathrm{d}\alpha \tag{29}$$

In this method, the current variational lower bound expressed by (29) is maximized consecutively. To explain more in detail, let us consider a set of observations $\{\vec{X}_1, \ldots, \vec{X}_{(t-1)}\}$. Then, a new observation $\vec{X}_t$ arrives and we maximize and update the current lower bound $\mathcal{L}^{(t)}(Q)$ corresponding to $Q(\vec{Z}_t)$, while $Q(\vec{\alpha})$ and π_j is set to $Q^{t-1}(\vec{\alpha})$ and π_j^{t-1}, respectively. Thus, the variational solution to $Q(\vec{Z}_t)$ is as follows:

$$Q(\vec{Z}_t) = \prod_{j=1}^{M} r_{tj}^{Z_{tj}} \tag{30}$$

$$r_{ij} = \frac{\tilde{r}_{tj}}{\sum_{j=1}^{M} \tilde{r}_{tj}} \tag{31}$$

$$\tilde{r}_{tj} = \exp \left\{ \ln \pi_j^{(t-1)} + \tilde{R}_j + \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} - 1 \right) \ln x_{tl} - \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} + 1 \right) \ln(1 - x_{tl}) - \mid \overline{\alpha}_j \mid \ln \left[1 + \sum_{l=1}^{D} \frac{x_{tl}}{(1 - x_{tl})} \right] \right\} \tag{32}$$

$\tilde{R}_j$ is calculated in Appendix A. Then, with the application of the gradient method, we set $Q(\vec{Z}_t)$ fixed, so that the lower bound is maximized corresponding to $Q^{(t)}(\vec{\alpha})$ and $\pi_j^{(t)}$. Therefore, the natural gradients are estimated by multiplying the gradients of the parameters with the inverse of the coefficient matrix, which is then removed so that the natural gradients for the posterior probabilities can be computed for an efficient online learning framework. Thus, we have the optimal solutions for parameters'

updates:

$$\mathcal{Q}^{(t)}(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{d=1}^{D}\mathcal{G}\left(\alpha_{jd}^{(t)} \mid u_{jd}^{*(t)}, v_{jd}^{*(t)}\right) \tag{33}$$

$$u_{jd}^{*(t)} = u_{jd}^{*(t-1)} + \rho_t \Delta u_{jd}^{*(t)} \tag{34}$$

$$v_{jd}^{*(t)} = v_{jd}^{*(t-1)} + \rho_t \Delta v_{jd}^{*(t)} \tag{35}$$

The solution for the mixing coefficient $\pi_j^{(t)}$ is:

$$\pi_j^{(t)} = \pi_j^{(t-1)} + \rho_t \Delta \pi_j^{(t)} \tag{36}$$

where ρ_t denotes the learning rate [53] described by (37) with two constraints, $\epsilon \in (0.5, 1]$ and $\eta \geq 0$.

$$\rho_t = (\eta_0 + t)^{-\epsilon} \tag{37}$$

The main idea of the learning rate is to ignore the previous wrong estimations of the lower bound and accelerate the convergence rate. Therefore, the natural gradients are as follows:

$$\begin{aligned}\Delta u_{jl}^{*(t)} &= u_{jl}^{*(t)} - u_{jl}^{*(t-1)} = \\ &u_{jl} - +Nr_{tj}\overline{\alpha}_{jl}\left[\psi\left(\sum_{l=1}^{D}\overline{\alpha}_{jl}\right) - \psi\left(\overline{\alpha}_{jl}\right)\right. \\ &\left. + \sum_{d\neq s}^{D}\psi'\left(\sum_{l=1}^{D}\overline{\alpha}_{jl}\right) \times \overline{\alpha}_{js}\left(\left\langle \ln \alpha_{js}\right\rangle - \ln \overline{\alpha}_{js}\right)\right]\end{aligned} \tag{38}$$

$$\begin{aligned}\Delta v_{jl}^{*(t)} &= v_{jl}^{*(t)} - v_{jl}^{*(t-1)} = v_{jl} \\ &- Nr_{tj}\left[\ln x_{td} - \ln(1 - x_{tl}) - \ln\left[1 + \sum_{l=1}^{D}\frac{x_{tl}}{(1 - x_{tl})}\right]\right]\end{aligned} \tag{39}$$

$$\Delta \pi_j^{(t)} = \pi_j^{(t)} - \pi_j^{(t-1)} = \left(\frac{N}{t}\right) r_{tj} - \pi_j^{(t-1)} \tag{40}$$

where $\left\langle \ln \alpha_{jd}\right\rangle$ and $< (\ln \alpha_{jd} - \ln \overline{\alpha}_{jd})^2 >$ are similar to (25) and (26), respectively. When a new data point arrives, an additional distribution is added to the lower bound.

There are two constraints expressed by (41) that ensure the convergence of lower bound as the online learning framework can be considered as a stochastic approximation:

$$\sum_{t=1}^{\infty}\rho_t = \infty, \qquad \sum_{t=1}^{\infty}{\rho_t}^2 < \infty \tag{41}$$

Our model is described completely in Algorithm 2. We applied k-means to initialize the parameter. Consequently, the variational solutions are updated by iteration until convergence. The clusters with minimal weight close to zero are automatically removed.

ALGORITHM 2 Online variational framework for MBMM

1. Choose a large initial number of components M.
2. Initialize values for u_{jd} and v_{jd}.
3. Initialize the value of r_{ij} using K-Means algorithm.
4. For $t = 1$ to N
5. **repeat**
6. The variational E-step:
7. Estimate the expected values.
8. Calculate learning rate by (37).
9. Compute $\Delta u_{jd}^{*(t)}$, $\Delta v_{jd}^{*(t)}$ and $\Delta \pi_j^{(t)}$ by (38) to (40).
10. The variational M-step:
11. Update the variational solutions to update $Q(\mathcal{Z}_t)$
12. Update the variational solutions to update $Q(\vec{\alpha})$.
13. **until** Convergence criterion is reached.

5 | EXPERIMENTAL RESULTS

In this section, we validate the performance of online variational learning of multivariate Beta mixture model (OVMBMM) on four strong candidates in real-world medical applications, namely, image segmentation of colorectal cancer, multi-class colon tissue analysis, digital imaging in skin lesion diagnosis and computer aid detection (CAD) of Malaria. It is noteworthy to mention that our main motivation to focus on medical applications is that advanced analytical and statistical methods provide more precise information to healthcare systems which is a valuable asset for the patient care as having more information, better understanding and improved analysis results in making proper decisions in different steps such as screening, diagnosis and treatment. The significance of machine learning in healthcare applications is enhanced specially in development of high-performance medical image processing systems. Computer-aided detection (CADe) detects clinically significant objects from medical images and computer-aided diagnosis (CADx) generally confronts with processing and analysing high dimensional datasets which is beyond the scope of human capability. It is obvious that in both of these domains, advanced clinical insights ultimately lead to improve quality of services, better outcomes, lower healthcare costs, and increased patient satisfaction. In some disciplines such as radiology and pathology, identification of abnormalities and marking the critical areas are vital to improve efficiency, reliability, and accuracy of diagnosis. Moreover, such medical testing techniques generate large scale datasets for which online variational inference is a proper modelling method. Here, we compare four algorithms, namely, batch variational multivariate Beta mixture model (BVMBMM), online variational multivariate Beta mixture model (OVMBMM), batch variational Gaussian mixture model (BVGMM) and online variational Gaussian mixture model (OVGMM) in terms of their accuracy based on confusion matrix and Jaccard similarity index for image segmentation.

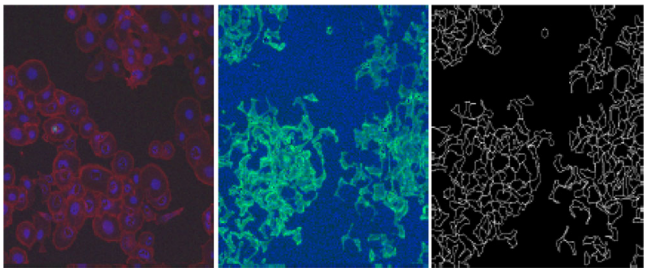

FIGURE 4 Sample images from colon tissues

5.1 | Image segmentation in colorectal cancer

According to World Health Organization (WHO) reports, cancer is the second leading cause of death globally, taking the life of 1 in 6 people, accounting for an estimated 9.6 million deaths in 2018 [54]. Colorectal cancer with 1.80 million cases, has the third place in the ranking of most common cancers and secondly ranked in most typical causes of cancer death with 862,000 deaths. Early detection and treatment has a great impact on reducing cancer mortality. By early identification and avoiding delays in care, the patient is more likely to survive by responding effectively to treatments. This goal is achieved by awareness, accessing clinical evaluation, diagnosis and having access to treatment [54]. One of the valuable solutions to avoid late stages detection is screening which aims to find individuals with abnormalities, pre-cancer and not developed symptoms. As one of the main steps of screening, tissue or cell samples can be taken from intestine or stomach for determining causes of abnormalities or presence and effects of cancer. Hence, histopathology analysis has a significant role and poses critical challenge as biological tissues have various structures and precise tumour segmentation, accurate pattern detection is a tough task for humans. In recent years, since tissue specimens were digitized, automated analysis of histopathology slides [55] has become a key requirement to asses quantitative morphology, cancer aggressiveness grading and reliable differentiation of various tumour types which is reflected by the formation and architecture of glands. Subsequently, machine learning techniques have demonstrated superior performance over conventional methods [56]. Here we focus on two applications related to colorectal cancers. First, image segmentation of a publicly available collection of microscopy images of colon cancer cells from broad bioimage benchmark collection (BBBC018v1) [57, 58]. The image set consists of 56 fields of view (four from each of 14 samples). Because there are three channels, there are 168 image files. The samples were stained with Hoechst 33342, pH3, and phalloidin. Hoechst 33342 is a DNA stain that labels the nucleus. Phospho-histone H3 indicates mitosis. Phalloidin labels actin, which is present in the cytoplasm. This image set is accompanied by a set of ground truth data to test automated image analysis against them. The ground truth set consists of outlines of nuclei and cells. In Figure 4, some examples of tissues and nucleus with their corresponding ground truth are illustrated.

TABLE 1 Accuracy of image segmentation for colon cancer dataset

Method	Accuracy (%)
OVMBMM	95.77
OVGMM	87.41
BVMBMM	90.31
BVGMM	84.33

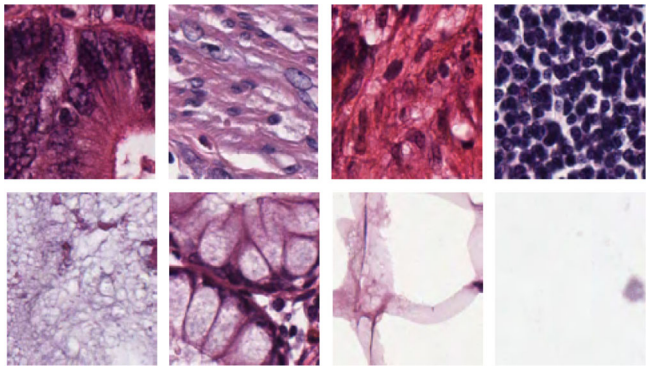

FIGURE 5 Sample images from colon dataset

The results of validating our proposed frameworks based on Jaccard similarity index are presented in Table 1 which prove that our model outperforms the three other alternatives.

5.2 | Multiclass colon tissue analysis

The second application is multiclass tissue clustering problem and categorization of a collection of textures in histological images of human colorectal cancer. The term texture refers to specific properties, pattern and structure of image regions. In medical image analysis, texture analysing methods are applied to classify tissue types. Human solid tumours are complex structures that typically several distinct tissue types are integrated in tumours consisting of non-malignant tissues, necrotic regions, tumour stroma, immune cell infiltration and islets of remaining. Moreover, tumour progression over time leads to changes in the architecture of tissue. In the digital pathology, automatic recognition of different tissue types assists to estimate the tumour/stroma ratio on histological samples and can provide quantitative and high-throughput analysis of the tumour tissue. In this paper to assess the performance, we evaluate our models by a collection of textures in colorectal cancer histology [59, 60] which is publicly available. It includes 5000 histological images of human colorectal cancer consisting of eight different types of tissue. In Figure 5, three samples of eight tissue classes are shown which enhance a variety of illumination, stain intensity and tissue textures. These classes are tumour epithelium, simple stroma that is homogeneous composition including tumour or extra-tumoural stroma, smooth muscle, single tumour or immune cells and/or single immune cell, complex stroma

TABLE 2 Accuracy of four models tested on multiclass colon tissue analysis

Method	Accuracy	Precision	Recall	F1-score
OVMBMM	93.16	93.24	92.32	93.2
OVGMM	85.38	85.6	85.4	85.49
BVMBMM	89.74	89.96	89.67	89.85
BVGMM	82.07	82.15	82.05	82.1

containing single tumour cells and/or few immune cells, immune cells including immune-cell conglomerates and sub-mucosal lymphoid follicles, debris including necrosis, haemorrhage and mucus, normal mucosal glands, adipose tissue and background without any tissue. As an important step, we extracted the feature of each image using one of the most popular techniques, namely, scale-invariant feature transform (SIFT) [61] and bag of visual words (BOVW). The general idea of this method is to represent an image as a set of features which include key points and descriptors. The keypoints of each image are invariant to geometrical transformation and illumination and descriptors are the description of these points which both are extracted by SIFT. Consequently, we construct vocabularies with key points and descriptors to represent each image as a frequency histogram of features which could be applied in image categorization to find images with similar pattern which could be differentiated by histopathological evaluation and the tissue composition could be quantified. The outputs of testing the performance of our algorithms are illustrated in Table 2 which obviously shows the superior performance of OVMBMM.

5.3 | Digital imaging in Melanoma lesion detection and segmentation

As stated by WHO, 1.04 million cases of skin cancer were reported in 2018 and it was ranked as the fifth common cancer [54]. The major cause of death from skin cancer is malignant melanoma which is caused by the abnormal multiplication of cells. However, it is far less prevalent than non-melanoma skin cancers. This type of cancer is primarily diagnosed visually. After initial clinical screening and dermoscopic analysis, a biopsy and histopathological sample is analysed. Digital imaging can help to recognize and treat in its earliest stages which lead to reduce melanoma mortality as it is readily curable. Automated diagnosis and digital images of skin lesions can aid in the diagnosis of melanoma through teledermatology. The standard quality of skin lesion imaging has a great impact on early detection and results in improvement of the efficiency, effectiveness, and accuracy of melanoma diagnosis. Nevertheless, unprofessional screening results in unnecessary biopsies and excisions of benign skin lesions. However, it is difficult to distinguish early-stage melanoma from benign skin lesions with similar structure which may lead to missing positive cases, useless clinical advanced examinations and misclassifying the benign and malignant melanoma. Thus, the expertise of the examiner and clinical setting have significant role. Evolution of digital imaging in skin lesion diagnosis permits the early detection of atypical lesions. Therefore, unnecessary biopsies of benign tumours are decreased or avoided. Recent enhancements in computer vision, machine learning algorithms and digital dermoscopic techniques can assist in image segmentation and retrieval, facilitate follow up and reduce unbiased diagnosis and misclassification rate therefore. These admirable advantages led to gaining the attention of researchers and increasing the focus towards computer aided systems in the last few decades. To evaluate automated Melanoma region segmentation using dermoscopic images, we tested our models, on a public dataset of ISIC [62], containing 23,906 images of skin lesions with their corresponding ground truths. In Figure 6, six samples of this dataset and their ground truth are illustrated. Similar to previous experiments, we compared four models. The outputs are presented in Table3 based on Jaccard similarity index. As it is shown, OVMBMM is more accurate than the other algorithms.

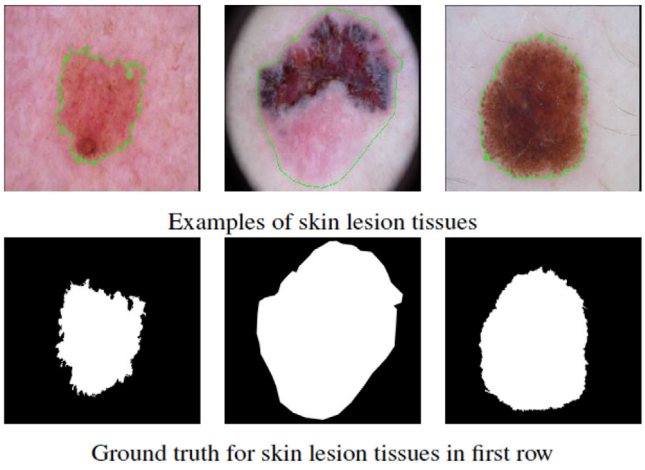

FIGURE 6 Sample images from skin dataset

TABLE 3 Accuracy of skin tumour segmentation

Method	Accuracy (%)
OVMBMM	94.29
OVGMM	87.53
BVMBMM	92.09
BVGMM	82.77

5.4 | Computer aid detection of Malaria

Malaria is a serious infectious disease caused by a blood parasite which is injected into the human body by female Anopheles mosquito. Considering the statistics announced by WHO, 219 million Malaria cases and 435,000 Malaria deaths were

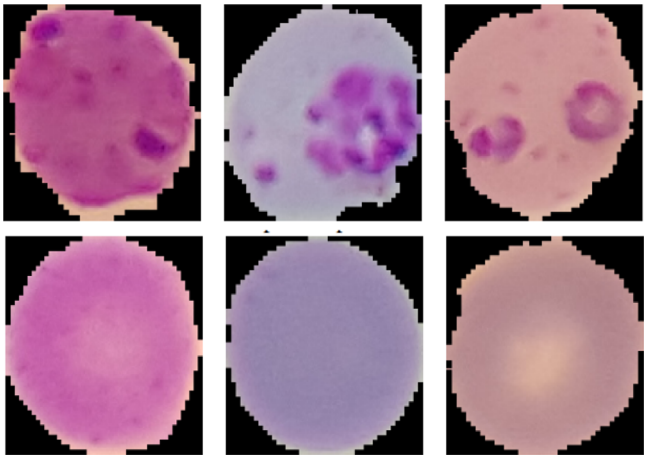

FIGURE 7 Sample images from Malaria dataset

globally reported in 2017 [63]. To manage and monitor this disease efficiently, it is crucial to diagnose it promptly and accurately as misdiagnosis can lead to significant morbidity and mortality. Therefore, with the help of parasitological and clinical microscopy which is considered as the mainstay of parasite-based diagnosis, the infection could be identified and confirmed precisely. The microscopy examination of Malaria, as the most prevalent and commonly practiced method, involves visual examination blood smears to test for the presence or absence of parasite in the blood and quantification of parasitemia, specie identification and life cycle classification. However, we should bear in mind that acceptable microscopy service with consistently accurate results is time consuming and costly and depends on the qualification of experts and load of samples. WHO reported that more than 208 million patients were tested by microscopic examination in 2017. Such massive number of ongoing examinations indicates the significance of process automation in analysis of samples. In order to overcome the issues such as error-prone and timely procedure, CAD and mathematical morphology are applied as effective tools for computer aided Malaria detection. These techniques are widely used for image processing purposes and employed successfully in biomedical image analysis. However, computer vision techniques for diagnosis, recognition and differentiation between non-parasitic and infected samples represent a relatively new domain of research. In our work, we applied our models on a dataset provided by NIH including thin blood smear slide images from the Malaria Screener research activity [64]. The dataset contains a total of 27,558 cell images with equal instances of parasitized and uninfected cells. A few examples of this dataset are illustrated in Figure 7 including six parasitized and six uninfected blood smear samples. In this experiment, the features are extracted by BOVW and SIFT. Finally to evaluate the performance of our method, we compared the results of four models which are illustrated in Table 4 indicating that OVMBMM has more accurate outputs. It is noteworthy to mention that these results clarify that online variational learning is a robust method as physicians are analysing large amount of pathological samples.

TABLE 4 Accuracy of four models tested on Malaria dataset

Method	Accuracy	Precision	Recall	F1-score
OVMBMM	92.5	90.47	95.11	92.68
OVGMM	82.5	77.08	92.5	84.09
BVMBMM	88.75	86.04	92.5	89.15
BVGMM	80.62	77.52	86.25	81.65

6 | CONCLUSION

This article introduces a novel unsupervised learning approach based on variational inference of finite multivariate Beta mixture model with the main focus on medical applications. Considering rich and various forms of medical information, artificial intelligence has a great impact on diagnosis and treatment of diseases. We developed our models based on variational Bayesian inference framework as a powerful alternative to deterministic methods such as maximum likelihood and conventional Bayesian inference, which has high computational cost. In our proposed method, convergence and simultaneously estimation of parameters and model complexity is guaranteed within an iterative process. Then, we employ online variational learning as an extension to classic method which keeps not only the advantages of previous models, but also speeds up the convergence rate significantly. Indeed, the online algorithm has a great capability to handle different demanding large scale datasets in real time. The performances of the proposed methods are validated with four challenging medical applications namely, image segmentation in colorectal cancer, multiclass tissue clustering, digital imaging in Melanoma lesion detection and segmentation and computer aid detection of Malaria. We compare four algorithms, batch variational multivariate Beta mixture model, online variational multivariate Beta mixture model, batch variational Gaussian mixture model and online variational Gaussian mixture model and evaluate their performance accuracy. According to the obtained results, we can clearly see that the OVMBMM outperforms the three other methods in terms of all four applications. Future works could be devoted to the extension of the proposed model within a non-parametric Bayesian framework to add more flexibility. Also, in our current model, we supposed that all extracted features have the same importance which is a limitation. Consequently, we will extend our study to a model which includes a feature selection strategy.

ACKNOWLEDGEMENT

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

ORCID

Narges Manouchehri https://orcid.org/0000-0002-3011-5162

REFERENCES

1. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science 349(6245), 255–260 (2015)
2. Zhang, Q., et al.: A survey on deep learning for big data. Inf. Fusion 42, 146–157 (2018)
3. Yan, L., Yoshua, B., Geoffrey, H.: Deep learning. Nature 521(7553), 436–444 (2015)
4. Yao, G., Lei, T., Zhong, J.: A review of convolutional-neural-network-based action recognition. Pattern Recognit. Lett. 118, 14–22 (2019)
5. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26 (2017)
6. Freire.Obregón, D., Narducci, F., Barra, S., Castrillón.Santana, M.: Deep learning for source camera identification on mobile devices. Pattern Recognit. Lett. 126, 86–91 (2019)
7. Hossain, M.S., Muhammad, G.: Emotion recognition using deep learning approach from audio–visual emotional big data. Inf. Fusion 49, 69–78 (2019)
8. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., et al.: Deep learning for generic object detection: A survey. Int. J. Comput. Vis. 128(2), 261–318 (2020)
9. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE Trans. Neural Netw. Learning Syst. 30(11), 3212–3232 (2019)
10. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. 27(1), 38–49 (2017)
11. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
12. Samadi, F., Akbarizadeh, G., Kaabi, H.: Change detection in sar images using deep belief network: a new training approach based on morphological images. IET Image Process. 13(12), 2255–2264 (2019)
13. Sharifzadeh, F., Akbarizadeh, G., Kavian, Y.S.: Ship classification in sar images using a new hybrid cnn–mlp classifier. J. Ind. Soc. Remote Sens. 47(4), 551–562 (2019)
14. Cheng, G., et al.: When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. IEEE Trans. Geosci. Remote Sens. 56(5), 2811–2821 (2018)
15. Khan, S., et al.: Energy-efficient deep cnn for smoke detection in foggy IoT environment. IEEE Internet of Things J. 6(6), 9237–9245 (2019)
16. Obinikpo, A.A., Kantarci, B.: Big sensed data meets deep learning for smarter health care in smart cities. J. Sensor Actuator Netw. 6(4), 26 (2017)
17. Kumar, M.A., Kumar, M., Sheshadri, H.: Computer aided detection of clustered microcalcification: A survey. Curr. Med. Imaging 15(2), 132–149 (2019)
18. Tresp, V., Overhage, J.M., Bundschus, M., Rabizadeh, S., Fasching, P.A., Yu, S.: Going digital: a survey on digitalization and large-scale data analytics in healthcare. Proc. IEEE 104(11), 2180–2206 (2016)
19. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 2(4), 230–243 (2017)
20. Bernal, J., Kushibar, K., Asfaw, D.S., Valverde, S., Oliver, A., Martí, R., et al.: Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artif. Intell. Med. 95, 64–81 (2019)
21. Tokunaga, H., et al.: Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12597–12606, IEEE, Piscataway, NJ (2019)
22. Bizopoulos, P., Koutsouris, D.: Deep learning in cardiology. IEEE Rev. Biomed. Eng. 12, 168–193 (2018)
23. Roy, K., Banik, D., Bhattacharjee, D., Nasipuri, M.: Patch-based system for classification of breast histology images using deep learning. Comput. Med. Imaging Graph. 71, 90–103 (2019)
24. Hegde, R.B., et al.: Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. Biocybernetics Biomed. Eng. 39(2), 382–392 (2019)
25. Yasaka, K., et al.: Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: A preliminary study. Radiology 286(3), 887–896 (2018)
26. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency 2, (2017)
27. Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput. Appl. 32, 18069–18083 (2019)
28. Miotto, R., et al.: Deep learning for healthcare: review, opportunities and challenges. Briefings Bioinform. 19(6), 1236–1246 (2018)
29. McLachlan, G.J., Lee, S.X., Rathnayake, S.I.: Finite mixture models. Annu. Rev. Stat. Appl. 6, 355–378 (2019)
30. Tian, F., Zhou, Q., Yang, C.: Gaussian mixture model-hidden markov model based nonlinear equalizer for optical fiber transmission. Opt. Express 28(7), 9728–9737 (2020)
31. Bouguila, N., Ziou, D.: Unsupervised selection of a finite Dirichlet mixture model: an mml-based approach. IEEE Trans. Knowl. Data Eng. 18(8), 993–1009 (2006)
32. Fan, W., Bouguila, N., Ziou, D.: Variational learning of finite Dirichlet mixture models using component splitting. Neurocomputing 129, 3–16 (2014)
33. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Trans. Pattern Anal. Mach. Intell. 29(10), 1716–1731 (2007)
34. Fan, W., Bouguila, N.: Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. Pattern Recognit. 46(10), 2754–2769 (2013)
35. Fan, W., Bouguila, N.: Online learning of a Dirichlet process mixture of beta-liouville distributions via variational inference. IEEE Trans. Neural Netw. Learning Syst. 24(11), 1850–1862 (2013)
36. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc.: Ser. B 39(1), 1–22 (1977)
37. Husmeier, D.: The Bayesian evidence scheme for regularizing probability-density estimating neural networks. Neural Comput. 12(11), 2685–2717 (2000)
38. Husmeier, D., Penny, W.D., Roberts, S.J.: An empirical evaluation of Bayesian sampling with hybrid monte carlo for training neural network classifiers. Neural Netw. 12(4-5), 677–705 (1999)
39. Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes (2013), arXiv:13016676
40. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24(3), 381–396 (2002)
41. Law, M.H., Figueiredo, M.A., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 26(9), 1154–1166 (2004)
42. Graham, M.W., Miller, D.J.: Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection. IEEE Trans. Signal Process. 54(4), 1289–1303 (2006)
43. Li, Y., Dong, M., Hua, J.: Simultaneous localized feature selection and model detection for gaussian mixtures. IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 953–960 (2008)
44. Sato, M.A.: Online model selection based on the variational Bayes. Neural Comput. 13(7), 1649–1681 (2001)
45. Olkin, I., Liu, R.: A bivariate beta distribution. Statistics Probability Lett. 62(4), 407–412 (2003)
46. Olkin, I., Trikalinos, T.A.: Constructions for a bivariate beta distribution. Statistics Probability Lett. 96, 54–60 (2015)
47. Bishop, C.M.: Pattern recognition and machine learning. New York: Springer 2006)
48. Lawrence, N.D., Bishop, C.M., Jordan, M.I.: Mixture representations for inference and learning in boltzmann machines. CoRR (2013), arXiv:13017393. http://arxiv.org/abs/1301.7393
49. Ma, Z., Leijon, A.: Bayesian estimation of beta mixture models with variational inference. IEEE Trans. Pattern Anal. Mach. Intell. 33(11), 2160–2173 (2011)

50. Blei, D.M., et al.: Variational inference for Dirichlet process mixtures. Bayesian Analysis 1(1), 121–143 (2006)
51. Fan, W., Bouguila, N., Ziou, D.: Variational learning for finite Dirichlet mixture models and applications. IEEE Trans. Neural Netw. Learning Syst. 23(5), 762–774 (2012)
52. Fan, W., Bouguila, N.: Online variational learning of generalized Dirichlet mixture models with feature selection. Neurocomputing 126, 166–179 (2014)
53. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems, pp. 856–864, MIT Press, Cambridge, MA (2010)
54. WHO Report. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 12 September 2018
55. Alagappan, M., et al.: Artificial intelligence in gastrointestinal endoscopy: The future is almost here. World J. Gastrointestinal Endoscopy 10(10), 239 (2018)
56. McCann, M.T., et al.: Automated histology analysis: Opportunities for signal processing. IEEE Signal Process. Mag. 32(1), 78–87 (2014)
57. Colon Dataset. https://data.broadinstitute.org/bbbc/BBBC018. Accessed: 10 February 2009
58. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature Methods 9(7), 637–637 (2012)
59. Zenodo Dataset. https://zenodo.org/record/53169.XvqPui2z3OQ. Accessed: 26 May 2016
60. Kather, J.N., et al.: Multi-class texture analysis in colorectal cancer histology. Scientific reports 6, 27988 (2016)
61. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)
62. ISIC Skin Dataset. https://www.isic-archive.com. Accessed: 4 May 2016
63. WHO Malaria Statistics. https://www.who.int/malaria/en/. Accessed: 30 November 2020
64. NIH Malaria Dataset. https://ceb.nlm.nih.gov/repositories/malaria-datasets/. Accessed: 14 August 2019

How to cite this article: Manouchehri N, Kalra M, Bouguila N. Online variational inference on finite multivariate Beta mixture models for medical applications. IET Image Process. 2021;15:1869–1882. https://doi.org/10.1049/ipr2.12154

APPENDIX A

Proof of equations $\mathcal{Q}(\mathcal{Z})$, $\mathcal{Q}(\vec{\alpha})$:

The variational solution $\mathcal{Q}_s(\Theta_s)$ is expressed by:

$$\ln \mathcal{Q}_s(\Theta_s) = \left\langle \ln p(\mathcal{X}, \Theta) \right\rangle_{t \neq s} + \text{const} \tag{A.1}$$

The additive constant term includes any term which is independent of $\mathcal{Q}_s(\Theta_s)$. $\mathcal{Q}(\mathcal{Z})$ and $\mathcal{Q}(\vec{\alpha})$ are derived from the logarithm of the joint distribution $p(\mathcal{X}, \Theta)$.

A.1 | Proof of Equation (16): variational solution of $\mathcal{Q}(\mathcal{Z})$

$$\ln \mathcal{Q}(Z_{ij}) = Z_{ij} \left[\ln \pi_j + R_j + \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} - 1 \right), \ln x_{il} \right. \tag{A.2}$$

$$- \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} + 1 \right) \ln(1 - x_{il})$$

$$\left. - \mid \overline{\alpha}_j \mid \ln \left[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})} \right] \right] + \text{const}$$

where,

$$R_j = \left\langle \ln \frac{\Gamma\left(\sum_{l=1}^{D} \alpha_{jl} \right)}{\prod_{l=1}^{D} \Gamma\left(\alpha_{jl} \right)} \right\rangle_{\alpha_{j1}, \ldots, \alpha_{jD}} \tag{A.3}$$

$$\left\langle \alpha_{jl} \right\rangle = \frac{u_{jl}}{v_{jl}} \tag{A.4}$$

As R_j is intractable and has not a closed form and standard variational inference can be applied indirectly. Thus, we approximate the lower bound to obtain a closed-form expression by the second-order Taylor series expansion. The function R_j is approximated about $\vec{\alpha}$. $\tilde{R}_j$ and $(\overline{\alpha}_{j1}, \ldots, \overline{\alpha}_{jD})$ are notations for approximation of R_j and $\vec{\alpha}$, respectively. The approximation of R_j is proved in [32] and after replacing it by $\tilde{R}_j$, optimization of (A.2) is tractable. So, the optimal solution for $\mathcal{Z}$ can be derived by:

$$\ln \mathcal{Q}(\mathcal{Z}) = \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} \ln \tilde{r}_{ij} + \text{const} \tag{A.5}$$

$$\ln \tilde{r}_{ij} = \ln \pi_j + \tilde{R} + \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} - 1 \right) \ln x_{il} \tag{A.6}$$

$$- \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} + 1 \right) \ln(1 - x_{il})$$

$$- \mid \overline{\alpha}_j \mid \ln \left[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})} \right] + \text{const}$$

By taking the exponential of both sides of (A.5), we will have:

$$\mathcal{Q}(\mathcal{Z}) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \tilde{r}_{ij}^{Z_{ij}} \tag{A.7}$$

By normalizing the distribution, $\mathcal{Q}(\mathcal{Z})$ is as follows where r_{ij} are positive and sum to one.

$$\mathcal{Q}(\mathcal{Z}) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{A.8}$$

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M} \tilde{r}_{ij}} \tag{A.9}$$

Thus, the standard result for $Q(\mathcal{Z})$ is:

$$\langle Z_{ij} \rangle = r_{ij} \tag{A.10}$$

A.2 | Proof of equation (17): variational solution of $Q(\vec{\alpha})$

Considering the assumption that the parameters α_{jl} are independent, $Q(\vec{\alpha})$ can be factorized as:

$$Q(\vec{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{M} Q(\alpha_{ij}) \tag{A.11}$$

Considering a specific factor $Q(\alpha_{ij})$, the variational optimization is derived by taking logarithm of the optimized factor given by: As in the other two cases the logarithm of the variational solution $Q(\alpha_{jl})$ is given by,

$$\begin{aligned} \ln Q(\alpha_{js}) &= \langle \ln p(\mathcal{X}, \Theta) \rangle_{\Theta \neq \alpha_{js}} \\ &= \sum_{i=1}^{N} \langle Z_{ij} \rangle \mathcal{J}(\alpha_{js}) + \alpha_{js} \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\ln x_{is} \right. \\ &\quad \left. - \ln(1 - x_{is}) - \ln \left[1 + \sum_{l=1}^{D} \frac{x_{is}}{(1 - x_{is})} \right) \right] \\ &\quad + (u_{js} - 1) \ln \alpha_{js} - \nu_{js} \alpha_{js} + \text{const} \end{aligned} \tag{A.12}$$

where,

$$\mathcal{J}(\alpha_{js}) = \left\langle \ln \frac{\Gamma\left(\alpha_s + \sum_{s \neq l}^{D} \alpha_{jl}\right)}{\Gamma(\alpha_s) \prod_{s \neq l}^{D} \Gamma(\alpha_{jl})} \right\rangle_{\Theta \neq \alpha_{js}} \tag{A.13}$$

Similar to R_j, $\mathcal{J}(\alpha_{js})$ is intractable and we its lower bound by calculating the first-order Taylor expansion with respect to $\overline{\alpha}_{js}$ which is expressed by:

$$\begin{aligned} \mathcal{J}(\alpha_{js}) \geq \overline{\alpha}_{js} \ln \alpha_{js} &\left[\psi\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right), -\psi(\overline{\alpha}_{js}), + \sum_{s \neq l}^{D} \overline{\alpha}_{jl} \right. \\ &\left. \times \psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{js} \right), \left(\langle \ln \alpha_{jl} \rangle - \ln \overline{\alpha}_{jl} \right) \right] + \text{const} \end{aligned} \tag{A.14}$$

This approximation is also found to be a strict lower bound of $\mathcal{J}(\alpha_{jl})$ and,

$$\begin{aligned} \ln \mathcal{Q}(\alpha_{js}) = \sum_{i=1}^{N} \langle Z_{ij} \rangle \overline{\alpha}_{js} \ln \alpha_{js} &\left[\psi\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right), -\psi(\overline{\alpha}_{js}) \right. \\ &\left. + \sum_{s \neq l}^{D} \overline{\alpha}_{js} \psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right), \left(\langle \ln \alpha_{jl} \rangle - \ln \overline{\alpha}_{jl} \right) \right] \end{aligned} \tag{A.15}$$

$$\begin{aligned} &+ \alpha_{js} \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\ln x_{is} - \ln(1 - x_{is}) \right. \\ &\left. - \ln \left[1 + \sum_{l=1}^{D} \frac{x_{is}}{(1 - x_{is})} \right] \right] \\ &+ (u_{jl} - 1) \ln \alpha_{jl} - \nu_{jl} \alpha_{jl} + \text{const} \\ &= \ln \alpha_{js} (u_{js} + \varphi_{js} - 1) - \alpha_{js} (\nu_{js} - \vartheta_{js}) + \text{const} \end{aligned}$$

where,

$$\begin{aligned} \varphi_{js} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \overline{\alpha}_{js} &\left[\psi\left(\sum_{l=1}^{D} \overline{\alpha}_{js} \right), -\psi(\overline{\alpha}_{js}) \right. \\ &\left. + \sum_{s \neq l}^{D} \overline{\alpha}_{jl} \psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right), \left(\langle \ln \alpha_{jl} \rangle - \ln \overline{\alpha}_{jl} \right) \right] \end{aligned} \tag{A.16}$$

$$\vartheta_{jl} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\ln x_{is} - \ln(1 - x_{is}) - \ln \left[1 + \sum_{l=1}^{D} \frac{x_{is}}{(1 - x_{is})} \right] \right] \tag{A.17}$$

Equation (A.15) is the logarithmic form of a Gamma distribution. By taking exponential of both the sides, we have:

$$Q(\alpha_{jl}) \propto \alpha_{jl}^{u_{jl} + \varphi_{jl} - 1} e^{-(\nu_{jl} - \vartheta_{jl}) \alpha_{jl}} \tag{A.18}$$

Thus, the optimal solution for the hyper-parameters u_{js} and ν_{js} given by:

$$u_{js}^{*} = u_{js} + \varphi_{jl}, \quad \nu_{js}^{*} = \nu_{js} - \vartheta_{js} \tag{A.19}$$

A.3 | Calculation of $\tilde{R}_j$ for equations (19) and (32)

$$\begin{aligned} \tilde{R}_j &= \ln \frac{\Gamma\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right)}{\prod_{l=1}^{D} \Gamma(\overline{\alpha}_{jl})} \\ &+ \sum_{l=1}^{D} \overline{\alpha}_{jl} \left[\psi\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right) - \psi(\overline{\alpha}_{jl}) \right] \times \left[\langle \ln \alpha_{jl} \rangle - \ln \overline{\alpha}_{jl} \right] \\ &+ \frac{1}{2} \sum_{l=1}^{D} \overline{\alpha}_{jl}^{2} \left[\psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right) - \psi'(\overline{\alpha}_{jl}) \right] \\ &\times \left\langle \left(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl} \right)^{2} \right\rangle \\ &+ \frac{1}{2} \sum_{a=1}^{D} \sum_{b=1, a \neq b}^{D} \overline{\alpha}_{ja} \overline{\alpha}_{jb} \left[\psi'\left(\sum_{l=1}^{D} \overline{\alpha}_{jl} \right) \right. \\ &\left. \times \left(\langle \ln \overline{\alpha}_{ja} \rangle - \ln \overline{\alpha}_{ja} \right), \times \left(\langle \ln \overline{\alpha}_{jb} \rangle - \ln \overline{\alpha}_{jb} \right) \right] \end{aligned} \tag{A.20}$$

A.4 | Lower bound $\mathcal{L}(Q)$ of (28)

$$
\begin{aligned}
\mathcal{L}(Q) &= \sum_{\mathcal{Z}} \int Q\left(\mathcal{Z}, \vec{\alpha}\right) \ln\left(\frac{p\left(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}\right)}{Q\left(\mathcal{Z}, \vec{\alpha}\right)}\right) \mathrm{d}\vec{\alpha} \\
&= \left\langle \ln p\left(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}\right)\right\rangle + \left\langle \ln p\left(\mathcal{Z} \mid \vec{\pi}\right)\right\rangle + \left\langle \ln p\left(\vec{\alpha}\right)\right\rangle - \\
&\langle \ln Q(\mathcal{Z})\rangle - \left\langle \ln Q\left(\vec{\alpha}\right)\right\rangle \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} r_{ij}\left[\ln \pi_j + \tilde{R}_j + \sum_{l=1}^{D}\left(\overline{\alpha}_{jl} - 1\right), \ln x_{il}\right. \\
&\quad - \sum_{l=1}^{D}\left(\overline{\alpha}_{jl} + 1\right)\ln(1 - x_{il}) \\
&\quad \left. - \mid \overline{\alpha}_j \mid \ln\left[1 + \sum_{l=1}^{D}\frac{x_{il}}{(1 - x_{il})}\right], -\ln r_{ij}\right] \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{u_{jl}\ln \nu_{jl} - \ln\Gamma(u_{jl}) + (u_{jl} - 1)\left\langle \ln \alpha_{jl}\right\rangle - \nu_{jl}\overline{\alpha}_{jl}\right\} \\
&\quad - \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{u^*_{jl}\ln \nu^*_{jl} - \ln\Gamma(u^*_{jl}) + (u^*_{jl} - 1)\left\langle \ln \alpha_{jl}\right\rangle - \nu^*_{jl}\overline{\alpha}_{jl}\right\}
\end{aligned}
\tag{A.21}
$$