# Variational Learning for Finite Shifted-Scaled Dirichlet Mixture Model and Its Applications

**Zeinab Arjmandiasl**

A Thesis
in
The Concordia Institute
for
Quality Systems Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Quality Systems
Engineering) at
Concordia University
Montréal, Québec, Canada

March 2020

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By : **Zeinab Arjmandiasl**

Entitled : **Variational Learning for Finite Shifted-Scaled Dirichlet Mixture Models and Its Applications**

and submitted in partial fulfillment of the requirements for the degree of

**Master of Applied Science in Quality Systems Engineering**

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee :

________________________________________ Chair
Dr. Fereshteh Mafakheri

________________________________________ External Examiner BCEE
Dr. Mohamed Ouf

________________________________________ Internal Examiner
Dr. Mohsen Ghafouri

________________________________________ Co-Supervisor
Dr. Nizar Bouguila

________________________________________ Co-Supervisor
Dr. Jamal Bentahar

Approved by Dr. Mohammad Mannan
Chair of Department or Graduate Program Director

Date: April-24-2020. ________________________________________

Dr. Amir Asif
Dean of Faculty of Engineering and Computer Science

# Abstract

## Variational Learning for Finite Shifted-Scaled Dirichlet Mixture Models and Its Applications

**Zeinab Arjmandiasl**

With the huge amount of data produced every day, the interest in data mining and machine learning techniques has been growing. Ongoing advancement of technology has made AI systems subject to different issues. Data clustering is an important aspect of data analysis which is the process of grouping similar observations in the same subset. Among known clustering techniques, finite mixture models have led to outstanding results that created an inspiration toward further exploration of various mixture models and applications. The main idea of this clustering technique is to fit a mixture of components generated from a predetermined probability distribution into the data through parameter approximation of the components. Therefore, choosing a proper distribution based on the type of the data is another crucial step in data analysis. Although the Gaussian distribution has been widely used with mixture models, the Dirichlet family of distributions have been known to achieve better results particularly when dealing with proportional and non-Gaussian data.

Another crucial part in statistical modeling is the learning process. Among the conventional estimation approaches, Maximum Likelihood (ML) is widely used due to its simplicity in terms of implementation but it has some drawbacks, too. Bayesian approach has overcome some of the disadvantages of ML approach via taking prior knowledge into account. However, it creates new issues such as need for additional estimation methods due to the intractability of parameters' marginal probabilities. In this thesis, these limitations are discussed and addressed via defining a variational learning framework for finite shifted-scaled Dirichlet mixture model. The motivation behind applying variational inference is that compared to conventional Bayesian approach, it is much less computationally costly. Furthermore, in this method, the optimal number of components is estimated along with the parameter approximation automatically and simultaneously while convergence is guaranteed. The performance of our model, in terms of accuracy of clustering, is validated on real world challenging medical applications, including image processing, namely, Malaria detection, breast cancer diagnosis and cardiovascular disease detection as well as text-based spam email detection. Finally, in order to evaluate the merits of our model effectiveness, it is compared with four others widely used methods.

# Acknowledgements

I would like to express my profound gratitude to my supervisors Prof. Nizar Bouguila, and Prof. Jamal Bentahar for their continuous guidance and motivation, during the course of my master program. I have been honored to work under their supervision, and I really appreciate their support.

I have been fortunate to work with Ms. Narges Manouchehri, who has always guide me with her vast knowledge and support. I would not be able to complete the research without her valuable advises.

I would like to thank Mr. Kamal Maanicshah who shared his knowledge and taken the time and effort in order to make sure I am in the right track.

I would like to extend my gratitude to my fellow lab-mates, Hieu, Yogesh and others for the good thoughts and motivation during my master's program.

Many Thanks to my parents and sisters for their love and constant motivation that inspired me to move forward and complete my study.

# Contents

# List of Figures

# List of Tables

Chapter 1

# Introduction

## 1.1 Introduction and Related Work

Data mining and dealing with large and complex data has become an important part of decision-making procedures in various research domains. Clustering as a powerful statistical approach has been effectively and extensively applied in finding hidden patterns within data [1], [2], [3], [4]. Among all clustering and unsupervised learning methods, finite mixture models specifically have shown remarkable success in various applications [5], [6], [7]. The principle idea of this clustering technique is to fit a mixture of components derived from a predetermined distribution to the data via parameter estimation of the components [8], [9], [10].

There are two conventional methods to learn finite mixture models, namely, deterministic and Bayesian approaches. Each of these methods has certain downsides [8]. For instance, deterministic approaches such as maximum likelihood estimation (MLE) do not produce decent approximation given a small dataset, plus overfitting, being sensitive to initialization and converging to local maxima instead of the global one [11], [12], [13]. Bayesian approaches [14], [15], [16] with the incorporation of prior knowledge can overcome the aforementioned problems through simulation techniques, but they have their own drawbacks [17] such as being computationally complex and time consuming specially for high dimensional data [18], [19], [20].

A proposed alternative to avoid such disadvantages is variational Bayesian approach. In this framework, parameters are modeled by assuming an approximation for their true posterior and minimizing the Kullback–Leibler

(KL) divergence between approximated and true posterior [15], [21]. Variational learning has shown higher performance while being less computationally expensive in comparison to traditional Bayesian approach [21], [22], [23]. Another advantage of variational framework is its capability of finding the correct number of clusters automatically along with parameter estimation process. This is outstanding when we consider the MLE approach alone fails to find the true number of clusters and it needs a model selection criterion such as minimum message length [24]. Furthermore, the variational inference approach has proven to be superior to ML in parameter estimation [25], [26], [27], [28], [29] [30].

In mixture models, choosing a proper distribution to represent data is a pivotal step in order to achieve outstanding results. Gaussian mixture model (GMM) has been center of interest in most of previous data analysis researches due to its simplicity of estimation procedures [31], [32]. However, GMMs are not always the best solution for any type of data specially for non-Gaussian ones. Recent researches have proven the prominent capability of other distributions such as Dirichlet distributions over the well-known GMM in many applications particularly when dealing with proportional data [33], [34], [24]. The aforementioned works which have been done using maximum likelihood (ML) within Expectation Maximization framework (EM) [35], [36], have shown the flexibility and effectiveness of Dirichlet distribution family [37], in particular, generalized versions of Dirichlet such as scaled Dirichlet distribution [38] and shifted-scaled Dirichlet distribution [39]. We propose variational learning of a mixture model based on shifted-scaled Dirichlet distribution. This distribution is a generalized version of Dirichlet distribution which has two extra parameters, scale and location to make it more flexible and capable to spread out. It should be recalled that scaled Dirichlet distribution has just one more parameter compared to Dirichlet which provides less flexibility compared to shifted scaled Dirichlet distribution.

## 1.2 Contributions

The main contributions of this thesis are as follows:

- ✓ **Proposing a variational framework for finite shifted-scaled Dirichlet mixture models:**

  Our approach proposes a variational framework for learning shifted-scaled Dirichlet mixture models, which are developed based on a more generalized distribution than Dirichlet. Having two extra parameters, scale and location, makes this distribution more capable to spread out. This is done through finding a tractable lower bound on marginal likelihood by replacing the intractable parameter distribution with an approximated distribution. Our approach estimates the model parameters and detects the number of components automatically as part of the variational inference procedure. Therefore, it is computationally less expensive, and converges faster compared to conventional methods, in which the number of clusters is solved using a model selection criterion which itself requires validation.

- ✓ **Demonstrating the application of the proposed statistical models:**

  We evaluate the effectiveness of our proposed approach in parameter estimation and model selection on challenging applications. The first three applications are based on well-known medical datasets that could play an important role in early or fast diagnosis of diseases considering massive data gathered in medical sectors. The forth real application we have tested our model on, is spam detection application which has attracted lots of attention in information system security field. Furthermore, we compared the performance of our model with four other models including one deterministic model, maximum likelihood learning of Gaussian mixture models, and three variational models, namely, variational learning of Gaussian mixture models, variational learning of Dirichlet mixture models and variational learning of scaled Dirichlet mixture models. The result confirms the outperformance of variational shifted-scaled Dirichlet mixture models over the others in terms of overall accuracy in modeling real world data.

## 1.3 Thesis Overview

This thesis is organized as follows:

□ Chapter1 briefly introduces the fundamentals of data clustering using mixture models as well as mentioning their challenges. It also explains the motivation for the considered probability distribution and the variational learning approach.

□ Chapter 2 proposes a variational framework for shifted-scaled Dirichlet mixture models, which could simultaneously estimate the model parameters and determine the optimal number of components.

□ Chapter 3 presents the experimental results of the proposed approach on three medical real world applications, namely, Malaria detection, breast cancer diagnosis, and cardiovascular diseases detection as well as a demanding text application of spam detection.

□ Chapter 4 concludes our contribution and points out some limitations and some remarks for prospective future researches.

Chapter 2

# Proposed Statistical Framework

## 2.1 Model Specification

In this chapter, we first present shifted-scaled Dirichlet distribution. Afterward, the construction of mixture model based on this distribution will be explained.

### 2.1.1 Shifted-Scaled Dirichlet Distribution

A generalized version of Dirichlet distribution studied from a probabilistic point of view is shifted-scaled Dirichlet distribution. This random composition is derived by applying two operations in the simplex, perturbation and powering. A vector-space structure is defined by these operations which play the same role as the sum and product by scalars in real space [40]. This added set of parameters has been shown to attain many functional probability models [41] which can be employed to model compositional multivariate data.
Let us assume an observation, generated from a shifted scaled Dirichlet distribution (SSD), which is defined by $\vec{X} = (X_1, \ldots, X_D)$ as a random vector of porportional data where $\sum_{d=1}^{D} X_d = 1$, $0 \leq \ X_d \leq 1$. The parameters of this distribution are $\vec{\alpha} = \big(\alpha_1, \ldots, \alpha_D\big) \in \mathbb{R}_{+}^{D}$ , $\vec{\beta} = \big(\beta_1, \ldots, \beta_D\big) \in \mathbb{S}^{D}$ and

$\tau \in \mathbb{R}_+$. This distribution is expressed as follows:

$$p(\vec{X} \mid \theta) = \frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \frac{1}{\tau^{D-1}} \frac{\prod_{d=1}^{D} \beta_d^{-\dfrac{\alpha_d}{\tau}} X_d^{(\dfrac{\alpha_d}{\tau}-1)}}{\left(\sum_{d=1}^{D} (\dfrac{X_d}{\beta_d})^{\dfrac{1}{\tau}}\right)^{\alpha+}} \tag{2.1}$$

where $\Gamma(.)$ denotes the gamma function, $\vec{\alpha}$ is the shape parameter which represents the form of the distribution and $\alpha_+ = \sum_{d=1}^{D} \alpha_d$. $\vec{\beta}$ is the location parameter which refers to the data densities location and $\tau$ is a real scalar which tunes the variance of the density plot [39]. These parameters make our probability distribution remarkably flexible which empowers our model to fit various kinds of datasets.

### 2.1.2 Finite Shifted-Scaled Dirichlet Mixture Model

A convex combination of two or more probability distributions is called finite mixture model. Consider a set of $N$ independent identically distributed observations described by $\mathcal{X} = \left(\vec{X}_1, \ldots, \vec{X}_N\right)$ in which each sample is a $D$-dimensional vector, $\vec{X}_i = (X_{i1}, \ldots, X_{iD})$ assumed to be generated from Equation (2.2). We assume that this dataset could be explained by a finite mixture model including $M$ components as follows [42]:

$$p\left(\vec{X}_i \mid \Theta\right) = \sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \tag{2.2}$$

where $\pi_j$ is the mixing coefficient of component $j$ satisfying two constraints $0 < \pi_j < 1$, $\sum_{j=1}^{M} \pi_j = 1$.
$\Theta = \{\pi_1, \ldots, \pi_M, \vec{\theta}_1, \ldots, \vec{\theta}_M\}$ denotes the complete set of model parameters in which $\vec{\theta}_j = \{\vec{\alpha}_j, \vec{\beta}_j, \vec{\tau}_j\}$ represents the parameter vector for $j$th component. Therefore the likelihood function of SSD mixture model is given by:

$$p(\mathcal{X} \mid \vec{\pi}, \vec{\theta}) = \prod_{i=1}^{N} \left\{ \sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \right\} \tag{2.3}$$

For each $\vec{X}_i$, we introduce a $M$-dimensional random vector $\vec{Z}_i = \big(Z_{i1}, \ldots, Z_{iM}\big)$ where $Z_{ij} \in \{0, 1\}, \sum_{j=1}^{M} Z_{ij} = 1$. This latent variable is not directly observed in the model, but from which we infer the cluster to which $\vec{X}_i$ is assigned such that $Z_{ij} = 1$ if it belongs to cluster $j$ and 0 otherwise.
Therefore, the conditional probability distribution for the $N$ hidden variables $\mathcal{Z} = \big(\vec{Z}_1, \ldots, \vec{Z}_N\big)$ given $\vec{\pi}$ is defined as:

$$p(\mathcal{Z} \mid \vec{\pi}) = \prod_{j=1}^{N} \prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{2.4}$$

Thus, the conditional probability of data set $\mathcal{X}$ given the class labels $\mathcal{Z}$ is as follow where $\vec{\alpha} = (\vec{\alpha}_1, \ldots, \vec{\alpha}_M)$, $\vec{\beta} = (\vec{\beta}_1, \ldots, \vec{\beta}_M)$ and $\vec{\tau} = (\tau_1, \ldots, \tau_M)$:

$$p\big(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}, \vec{\beta}, \vec{\tau}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} \Big[p(\vec{X}_i \mid \vec{\theta}_j)\Big]^{Z_{ij}} \tag{2.5}$$

## 2.2 Variational Bayesian Learning

One of the crucial stages in fitting a dataset is learning the parameters of the mixture model that includes both parameter estimation and detection of number of components ($M$). In this section, we introduce variational Bayesian approach for learning of shifted-scaled Dirichlet mixture models (varSSDMM) which attain both of above-mentioned problems simultaneously.

### 2.2.1 Parameter Estimation

The joint distribution of all the random variables conditioned on $\vec{\pi}$ is given by:

$$p\left(\mathcal{X}, \Theta \mid \vec{\pi}\right) = p\left(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}, \vec{\beta}, \vec{\tau}\right) p(\mathcal{Z} \mid \vec{\pi}) p(\vec{\alpha}) p(\vec{\beta}) p(\vec{\tau}) \tag{2.6}$$

where $\Theta = \{\mathcal{Z}, \vec{\alpha}, \vec{\beta}, \vec{\tau}\}$ and the following conjugate priors is chosen for $\vec{\alpha}, \vec{\beta}, \vec{\tau}$, respectively:

$$p(\alpha_{jd}) = \mathcal{G}(\alpha_{jd} \mid u_{jd}, \nu_{jd}) = \frac{\nu_{jd}^{u_{jd}}}{\Gamma(u_{jd})} \alpha_{jd}^{u_{jd}-1} e^{-\nu_{jd}\alpha_{jd}} \tag{2.7}$$

$$p(\beta_{jd}) = \mathcal{D}(\beta_{jd} \mid \vec{h_j}) = \frac{\Gamma(\sum_{d=1}^{D} h_{jd})}{\prod_{d=1}^{D}\Gamma(h_{jd})}\prod_{d=1}^{D}\beta_{jd}^{h_{jd}-1} \tag{2.8}$$

$$p(\tau_j) = \mathcal{G}(\tau_j \mid q_j, s_j) = \frac{q_j^{s_j}}{\Gamma(q_j)}\tau_j^{q_j-1}e^{-s_j\tau_j} \tag{2.9}$$

$\{u_{jd}\}$, $\{\nu_{jd}\}$, $\{h_{jd}\}$, $\{q_{jd}\}$ and $\{s_{jd}\}$ are hyper-parameters which all satisfy the constraint of being greater than zero. $\mathcal{G}(.)$ and $\mathcal{D}(.)$ denote Gamma and Dirichlet distributions, respectively. Since the parameters are considered statistically independent, we can write:

$$p(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{d=1}^{D}p(\alpha_{jd}) \tag{2.10}$$

$$p(\vec{\beta}) = \prod_{j=1}^{M}\prod_{d=1}^{D}p(\beta_{jd}) \tag{2.11}$$

$$p(\vec{\tau}) = \prod_{j=1}^{M}p(\tau_j) \tag{2.12}$$

By substituting equations (2.10), (2.11), (2.12), (2.4) and (2.5) into the joint distribution defined in equation (2.6), we get:

$$p(\mathcal{X}, \Phi \mid \vec{\pi}) = \prod_{i=1}^{N}\prod_{j=1}^{M}(\pi_j\frac{\Gamma(\alpha_j+)}{\prod_{d=1}^{D}\Gamma(\alpha_{jd})}\frac{1}{\tau_j^{D-1}}\frac{\prod_{d=1}^{D}\beta_{jd}^{-\frac{\alpha_{jd}}{\tau_j}}\,x_{id}^{(\frac{\alpha_{jd}}{\tau_j}-1)}}{\left(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\right)^{\alpha_j+}})^{Z_{ij}} \tag{2.13}$$

$$\times\prod_{j=1}^{M}\prod_{l=1}^{D}[\frac{\nu_{jd}^{u_{jd}}}{\Gamma(u_{jd})}\alpha_{jd}^{u_{jd}-1}e^{-\nu_{jd}\alpha_{jd}} \times \frac{\Gamma(\sum_{d=1}^{D} h_{jd})}{\prod_{d=1}^{D}\Gamma(h_{jd})}\prod_{d=1}^{D}\beta_{jd}^{h_{jd}-1}]$$

$$\times\prod_{j=1}^{M}[\frac{q_j^{s_j}}{\Gamma(q_j)}\tau_j^{q_j-1}e^{-s_j\tau_j}]$$

A Graphical representation of this model is shown in Figure 2.1 where random variables are displayed within circles. Plates denote replication and the number of replications is shown in the lower right corner of it. The arrows represent the conditional dependencies among variables.

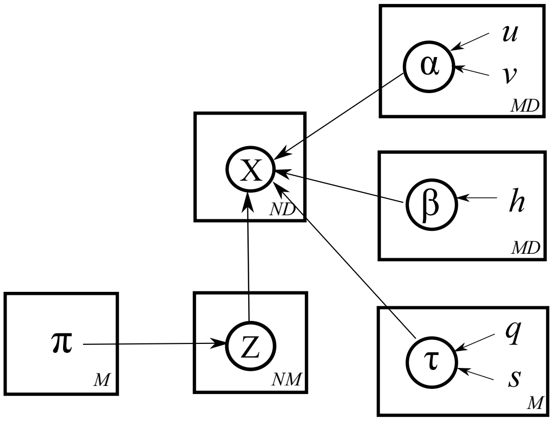


Figure 2.1: Graphical demonstration of the finite Shifted-Scaled Dirichlet mixture model

For learning our mixture model parameters and defining the correct number of components $M$ simultaneously, here we apply variational inference approach proposed in [27]. In this technique, a tractable lower bound $\mathcal{L}(Q)$ on the marginal likelihood $p\big(\mathcal{X} \mid \vec{\pi}\big)$ is used as below:

$$\ln p\big(\mathcal{X} \mid \vec{\pi}\big) = \ln \int p\big(\mathcal{X}, \Theta \mid \vec{\pi}\big) d\Theta = \\ \ln \int Q\big(\Theta\big) \frac{p\big(\mathcal{X}, \Theta \mid \vec{\pi}\big)}{Q\big(\Theta\big)} d\Theta \geq \int Q\big(\Theta\big) \ln \left( \frac{p\big(\mathcal{X}, \Theta \mid \vec{\pi}\big)}{Q\big(\Theta\big)} \right) d\Theta = \mathcal{L}\big(Q\big) \tag{2.14}$$

where $Q\big(\Theta\big)$ is introduced as an approximation for the true posterior $p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)$.
From equation (2.14), we find the following equation:

$$\mathcal{L}\big(Q\big) = \ln\ p\big(\mathcal{X} \mid \vec{\pi}\big) - KL\big(Q \mid\mid P\big) \tag{2.15}$$

where,

$$KL\big(Q \mid\mid P\big) = -\int Q\big(\Theta\big) \ln\left(\frac{p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)}{Q\big(\Theta\big)}\right) d\Theta \tag{2.16}$$

By maximizing the lower bound $\mathcal{L}(Q)$, the KL divergence reaches its minimum, zero, when $Q\big(\Theta\big) = p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)$. However, it is difficult to directly compute the true posterior for variational inference. Thus, $Q\big(\Theta\big)$ as a restricted family of distributions is taken into account.
We adopt an approximation method called mean-field theory [43], [44] which factorize $Q\big(\Theta\big)$ into tractable distributions of each parameter in the parameter space $\Theta$ as below:

$$Q\big(\Theta\big) = Q\big(\mathcal{Z}\big)Q\big(\vec{\alpha}\big)Q\big(\vec{\beta}\big)Q\big(\vec{\tau}\big) \tag{2.17}$$

Lower bound maximization is done through variational optimization of $\mathcal{L}(Q)$ with respect to each of the parameter distributions $Q_s\big(\Theta_s\big)$ which results in the following equation for a particular $Q_s\big(\Theta_s\big)$ [27]:

$$Q_s\big(\Theta_s\big) = \frac{exp\big\langle \ln\ p\big(\mathcal{X}, \Theta\big)\big\rangle_{j \neq s}}{\int exp\big\langle \ln\ p\big(\mathcal{X}, \Theta\big)\big\rangle_{j \neq s} d\Theta} \tag{2.18}$$

where $\big\langle \cdot \big\rangle_{j \neq s}$ indicates the expectation of all the distributions $Q_j\big(\Theta_j\big)$ excluding $j = s$. To apply the variational inference, all the parameter distributions $Q_j\big(\Theta_j\big)$ need to be initialized properly, since for optimal solution estimation $Q_s\big(\Theta_s\big)$ we loop over all the $Q_j\big(\Theta_j\big)$ except $j = s$. Afterward, each parameter get updated by an improved value which is calculated from equation (2.18) assuming the recent value of the other parameters altogether. Due to the convexity of the lower bound corresponding to each of the parameter distributions $Q_j\big(\Theta_j\big)$, convergence is certain [45], [46]. Finally, we get the optimal variational estimations for each Q in the parameter space $\Theta$ as follow (see Appendix A):

$$Q\big(\mathcal{Z}\big) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{2.19}$$

$$Q\big(\vec{\alpha}\big) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}\big(\alpha_{jd} \mid u_{jd}^{*}, \nu_{jd}^{*}\big) \tag{2.20}$$

$$Q(\vec{\beta}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{D}(\beta_{jd} \mid h^*_{jd}) \tag{2.21}$$

$$Q(\vec{\tau}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\tau_j \mid q^*_j, s^*_j) \tag{2.22}$$

where,

$$r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M} \rho_{ij}} \tag{2.23}$$

$$\begin{aligned} \rho_{ij} = exp\Bigg\{ & \ln \pi_j + \tilde{R}_j - (D-1) \ln \overline{\tau}_j \\ & + \sum_{d=1}^{D} \Bigg[ - \frac{\overline{\alpha}_{jd}}{\overline{\tau}_j} \ln \overline{\beta}_{jd} + (\frac{\overline{\alpha}_{jd}}{\overline{\tau}_j} - 1) \ln x_{id} \Bigg] \\ & - (\alpha_j +) \ln \Bigg( \sum_{d=1}^{D} (\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}} \Bigg) \Bigg\} \end{aligned} \tag{2.24}$$

$\tilde{R}_j$ is estimated as follows in which $\psi(.)$ and $\psi'(.)$ denotes digamma and trigamma functions, respectively (see Appendix A):

$$
\begin{aligned}
\tilde{R}_j = \ln & \frac{\Gamma\big(\sum_{d=1}^{D}\overline{\alpha}_{jd}\big)}{\prod_{d=1}^{D}\Gamma\big(\overline{\alpha}_{jd}\big)} \\
& + \sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg[\psi\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) - \psi\big(\overline{\alpha}_{jd}\big)\Bigg] \\
& \times \Big[\big\langle \ln\alpha_{jd}\big\rangle - \ln\overline{\alpha}_{jd}\Big] \\
& + \frac{1}{2}\sum_{d=1}^{D}\overline{\alpha}_{jd}^{2}\Bigg[\psi'\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) - \psi'\big(\overline{\alpha}_{jd}\big)\Bigg] \\
& \times \Big\langle\big(\ln\alpha_{jd} - \ln\overline{\alpha}_{jd}\big)^{2}\Big\rangle \\
& + \frac{1}{2}\sum_{a=1}^{D}\sum_{b=1,a\neq b}^{D}\overline{\alpha}_{ja}\,\overline{\alpha}_{jb}\Bigg\{\psi'\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) \\
& \times \Big(\big\langle \ln\overline{\alpha}_{ja}\big\rangle - \ln\overline{\alpha}_{ja}\Big) \times \Big(\big\langle \ln\overline{\alpha}_{jb}\big\rangle - \ln\overline{\alpha}_{jb}\Big)\Bigg\}
\end{aligned}
\tag{2.25}
$$

The hyperparameters $u_{jd}^{*}$, $\nu_{jd}^{*}$, $h_{jd}^{*}$, $q_{j}^{*}$ and $s_{j}^{*}$ are approximated as below (see Appendix A):

$$
u_{jd}^{*} = u_{jd} + \varphi_{jd} \qquad \nu_{jd}^{*} = \nu_{jd} + \vartheta_{jd}
\tag{2.26}
$$

$$
\begin{aligned}
\varphi_{jd} = \sum_{i=1}^{N}\big\langle Z_{ij}\big\rangle\overline{\alpha}_{jd}\Bigg[\psi\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) - \psi\big(\overline{\alpha}_{jd}\big) + \sum_{d\neq s}^{D}\psi'\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) \\
\times\,\overline{\alpha}_{js}\Big(\big\langle \ln\alpha_{js}\big\rangle - \ln\overline{\alpha}_{js}\Big)\Bigg]
\end{aligned}
\tag{2.27}
$$

$$
\vartheta_{jd} = \sum_{i=1}^{N}\big\langle Z_{ij}\big\rangle\left[\frac{1}{\tau_j}\ln\frac{\beta_{js}}{x_{is}} + \ln\Bigg(\sum_{d=1}^{D}\big(\frac{x_{id}}{\beta_{jd}}\big)^{\dfrac{1}{\tau_j}}\Bigg)\right]
\tag{2.28}
$$

$$
h_{jd}^{*} = h_{jd} + \kappa_{jd}
\tag{2.29}
$$

$$\kappa_{jd} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \Bigg[ \frac{-\bar{\alpha}_{js}}{\bar{\tau}_j} + \frac{\bar{\alpha}_{js}}{\bar{\tau}_j} \times (\frac{x_{is}}{\bar{\beta}_{js}})^{\frac{1}{\bar{\tau}_j}} \times \frac{1}{\sum_{d=1}^{D} (\frac{x_{id}}{\bar{\beta}_{jd}})^{\frac{1}{\bar{\tau}_j}}} \Bigg] \tag{2.30}$$

$$q_j^* = q_j + \delta_j \quad s_j^* = s_j - \varrho_j \tag{2.31}$$

$$\delta_j = \sum_{i=1}^{N} Z_{ij} \Bigg[ 1 - D + \frac{(\alpha_j +)}{\tau_j} \frac{\sum_{d=1}^{D} (\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}} \ln (\frac{x_{id}}{\beta_{jd}})}{\sum_{d=1}^{D} (\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}} \Bigg] \tag{2.32}$$

$$\varrho_j = \sum_{i=1}^{N} Z_{ij} \Bigg[ \sum_{d=1}^{D} \frac{\alpha_{jd}}{{\tau_j}^2} \ln (\frac{x_{id}}{\beta_{jd}}) \Bigg] \tag{2.33}$$

where the expected values in the preceding equations are as follow:

$$\big\langle Z_{ij} \big\rangle = r_{ij} \tag{2.34}$$

$$\overline{\alpha}_{jd} = \big\langle \alpha_{jd} \big\rangle = \frac{u_{jd}^*}{\nu_{jd}^*}, \qquad \big\langle \ln \alpha_{jd} \big\rangle = \psi\big(u_{jd}^*\big) - \ln \nu_{jd}^* \tag{2.35}$$

$$\Big\langle \big( \ln \alpha_{jd} - \ln \overline{\alpha}_{jd} \big)^2 \Big\rangle = \Big[ \psi\big(u_{jd}^*\big) - \ln u_{jd}^* \Big]^2 + \psi'\big(u_{jd}^*\big) \tag{2.36}$$

$$\overline{\beta}_{jd} = \big\langle \beta_{jd} \big\rangle = \frac{h_{jd}^*}{\sum_{d=1}^{D} h_{jd}^*} \tag{2.37}$$

$$\overline{\tau_j} = \big\langle \tau_j \big\rangle = \frac{q_j^*}{s_j^*} \tag{2.38}$$

### 2.2.2 Determination of the Number of Components

As mentioned before, $\vec{\pi}$ is considered as a parameter in variational learning. Thus, it is estimated via maximization of lower bond $\mathcal{L}(Q)$. The equation is obtained via setting the derivative of the lower bound with respect to $\vec{\pi}$ to zero:

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \tag{2.39}$$

In our experiments, the number of components has been initialized with large value such as 10 and with equal value of mixing coefficients. As the lower bound $\mathcal{L}(Q)$ is maximized in order to obtain the variational optimization of $Q\big(\mathcal{Z}\big)$, $Q\big(\vec{\alpha}\big)$, $Q\big(\vec{\beta}\big)$ and $Q\big(\vec{\tau}\big)$, the mixing coefficient $\vec{\pi}$ gets estimated as well. Therefore, the components which have trivial contribution to describe the data would have a close-to-zero mixing coefficients. Using automatic relevance determination [47], these components would be omitted from the model.

The steps of the variational algorithm for our model are described as follow:

**Algorithm 1** Variational learning algorithm of SSD

```
Initialize number of components $M$.
Initialize the hyper-parameters $\{u_{jd}\}$, $\{\nu_{jd}\}$, $\{h_{jd}\}$, $\{q_{jd}\}$ and $\{s_{jd}\}$.
Initialize $r_{ij}$ using *K-Means* algorithm.
**repeat**
    E-step of variational: update $Q\big(\mathcal{Z}\big)$, $Q\big(\vec{\alpha}\big)$, $Q\big(\vec{\beta}\big)$ and $Q\big(\vec{\tau}\big)$.
    M-step of variational: Maximize $\mathcal{L}(Q)$ with respect to recent value of $\vec{\pi}$ (2.39).
**until** Convergence criterion is reached.
Determine the number of components $M$ by omitting those with trivial mixing coefficients (smaller than $10^{-5}$).
Re-estimate new values of the parameters $(\mathcal{Z})$, $(\vec{\alpha})$, $(\vec{\beta})$, $(\vec{\tau})$ and $(\vec{\pi})$.
```

Chapter 3

# Experimental Result

## 3.1 Introduction

In this chapter, we evaluate the performance of our proposed model using three medical datasets, namely, Malaria, breast cancer and heart diseases as well as a challenging text dataset namely spam detection. The accuracy of the model mainly relies on the initialization of the hyperparameters including $\{u_{jd}\}$, $\{\nu_{jd}\}$, $\{h_{jd}\}$, $\{q_{jd}\}$ and $\{s_{jd}\}$. Thus, detecting a good set of initialized hyperparameters is an important step to obtain the optimal number of clusters and enhance the convergence rate. Besides, feature extraction and feature selection techniques are inevitable part of data pre-processing approach when dealing with image and text applications. Furthermore, scaling and normalization are crucial step and needs to be always considered for total performance improvement. To enhance the outperformance of our model, we compare it with four other models, namely, variational learning of scaled Dirichlet mixture model (varSDMM), variational learning of Dirichlet mixture model (varDMM), variational learning of Gaussian mixture model (varGMM) and maximum likelihood learning of Gaussian mixture model (GMM).

## 3.2 Malaria Detection

Malaria is a fatal disease in countries with tropical climates. It is caused by a parasite which is transmitted to humans through the bite of an infected mosquito. Based on the latest report released by WHO, in 2018, 405000 Malaria deaths has been registered among 228 million cases worldwide [48]. An accurate diagnosis is crucial in order to prevent from death and prevalence. Parasitological and clinical microscopy as a commonly used mean involves visual analysis of blood smears in order to detect the parasite in the blood as well as identifying the type, number and life cycle of the parasitemia. However, microscopy examination could be overwhelming and costly and very much relies on the qualification of the specialist and load of samples. The need for sample analysis automation has become undeniable considering the recent report published by WHO that 207 million suspected patients were tested via either an RDT or microscopy in 2018 [48]. We obtained a dataset from NIH containing slide images of blood smear released by the Malaria screener research activity [49]. The dataset has 27,558 images of blood cells with equal instances of infected and normal ones. Some samples of this dataset are shown in Figure 3.1 containing infected and normal blood smear instances. Acquiring a precise representation of the features of a dataset is an essential pre-processing task. In other words, an efficient descriptor containing most of the important features is needed. For this dataset we used Bag of visual words (BOVW) and SIFT [50] since it has well performed in various classification problems [45], [38], [51]. From the confusion matrix in Figure 3.2, we can see that 100% of the infected cells and 77.5% of the non-infected ones have been accurately detected which we can compare with other algorithms' confusion matrices shown in Figure 3.3. Finally, we compared the result of our model with four other models summarized in Table 3.1 denoting the outstanding accuracy of varSSDMM (88.7%). These results endorse the variational learning method on shifted-scaled Dirichlet as an effective approach.

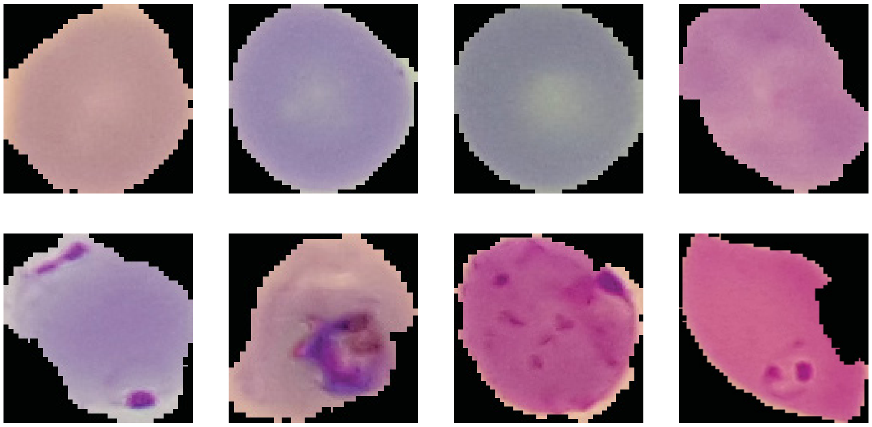
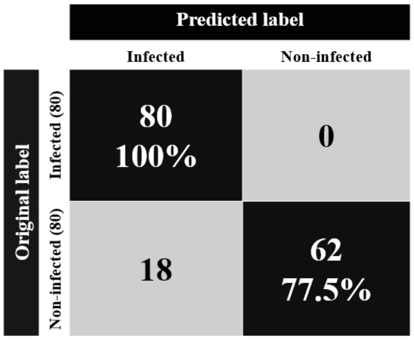


Figure 3.1: Malaria dataset

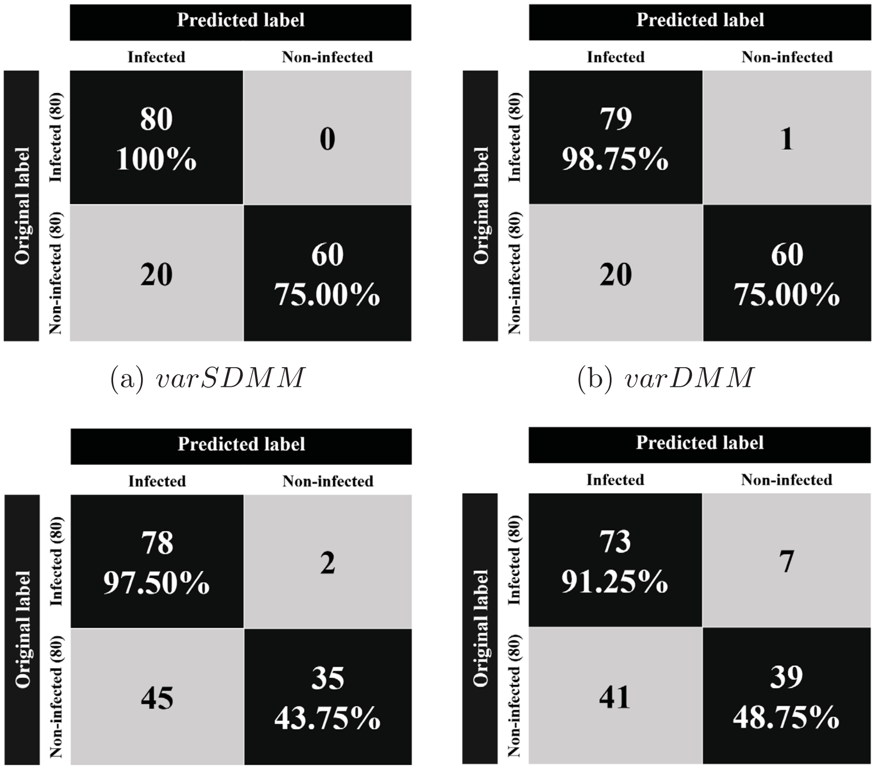


| | | Predicted label | |
|---|---|---|---|
| | | Infected | Non-infected |
| Original label | Infected (80) | 80<br>100% | 0 |
| | Non-infected (80) | 18 | 62<br>77.5% |

Figure 3.2: Malaria confusion matrix

Table 3.1: Model performance accuracy in malaria dataset

| Algorithm | varSSDMM | varSDMM | varDMM | varGMM | GMM |
|---|---|---|---|---|---|
| Accuracy(%) | 88.7 | 87.5 | 86.8 | 70.6 | 70.0 |



| Original label | Predicted label: Infected | Predicted label: Non-infected |
|---|---|---|
| Infected (80) | 80<br>100% | 0 |
| Non-infected (80) | 20 | 60<br>75.00% |

(a) *varSDMM*

| Original label | Predicted label: Infected | Predicted label: Non-infected |
|---|---|---|
| Infected (80) | 79<br>98.75% | 1 |
| Non-infected (80) | 20 | 60<br>75.00% |

(b) *varDMM*

| Original label | Predicted label: Infected | Predicted label: Non-infected |
|---|---|---|
| Infected (80) | 78<br>97.50% | 2 |
| Non-infected (80) | 45 | 35<br>43.75% |

(c) *varGMM*

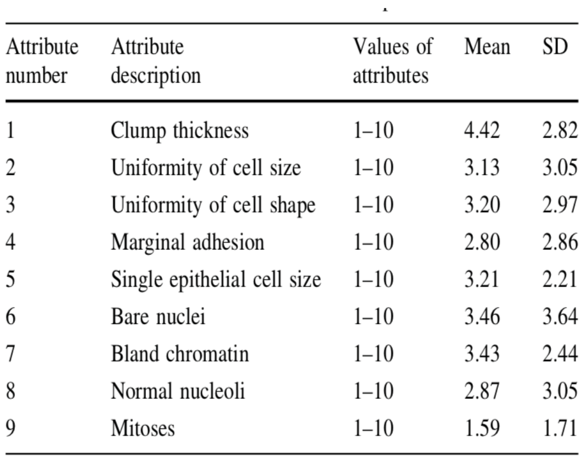
| Original label | Predicted label: Infected | Predicted label: Non-infected |
|---|---|---|
| Infected (80) | 73<br>91.25% | 7 |
| Non-infected (80) | 41 | 39<br>48.75% |

(d) *GMM*

Figure 3.3: Malaria confusion matrices for other algorithms

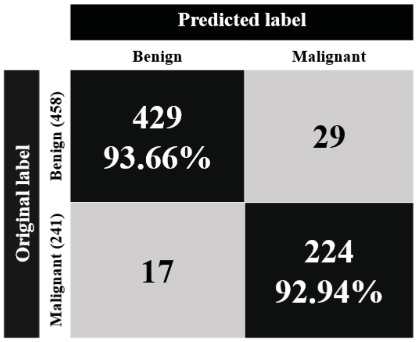
## 3.3 Breast Cancer Diagnosis

Breast cancer is the most prevalent type of cancers in women and the second common cancer worldwide, according to WCRF [52]. Early diagnosis and treatment play key roles to improve the survival rate of cancerous patients. Most of the breast cancer detections are done via screening imaging like sonography, mammography or MRI. Once a lump is detected, it is sampled for further analysis. Then a pathologist examines the tissue sample for detection of being benign or malignant. Among different ways of sampling, the fine needle aspiration (FNA) [53] is one of the standard and suitable means for medical diagnostic and decision-making processes. We evaluate our model over a publicly available breast cancer dataset named Wisconsin [54]. This dataset contains 699 samples with nine features that are computed from the images of breast lumps obtained via fine needle aspirate. In Figure 3.4, The attributes are graded from 1 to ten with 1 being the closest to benign and 10 the most anaplastic and malignant [55]. The mean and standard deviation of each attribute is detailed as well. It is noteworthy to mention that no single feature alone is enough to differentiate among benign and malignant instances and we employ all of them. There are (458) Benign and (241) Malignant cases in the set. The confusion matrix presented in Figure 3.5 clearly shows that the majority of the instances are correctly categorized, compare to other algorithms' confusion matrices shown in Figure 3.6. Lastly, the final result in Table 3.2 shows a significant improvement in clustering the benign and malignant samples using varSSDMM (93.4% accuracy) compared to the rest of algorithms. This result is achieved given that min-max scaling and normalization is applied on the dataset in order to improve the final result.



| Attribute number | Attribute description | Values of attributes | Mean | SD |
|---|---|---|---|---|
| 1 | Clump thickness | 1–10 | 4.42 | 2.82 |
| 2 | Uniformity of cell size | 1–10 | 3.13 | 3.05 |
| 3 | Uniformity of cell shape | 1–10 | 3.20 | 2.97 |
| 4 | Marginal adhesion | 1–10 | 2.80 | 2.86 |
| 5 | Single epithelial cell size | 1–10 | 3.21 | 2.21 |
| 6 | Bare nuclei | 1–10 | 3.46 | 3.64 |
| 7 | Bland chromatin | 1–10 | 3.43 | 2.44 |
| 8 | Normal nucleoli | 1–10 | 2.87 | 3.05 |
| 9 | Mitoses | 1–10 | 1.59 | 1.71 |

Figure 3.4: Attributes of Wisconsin Dataset

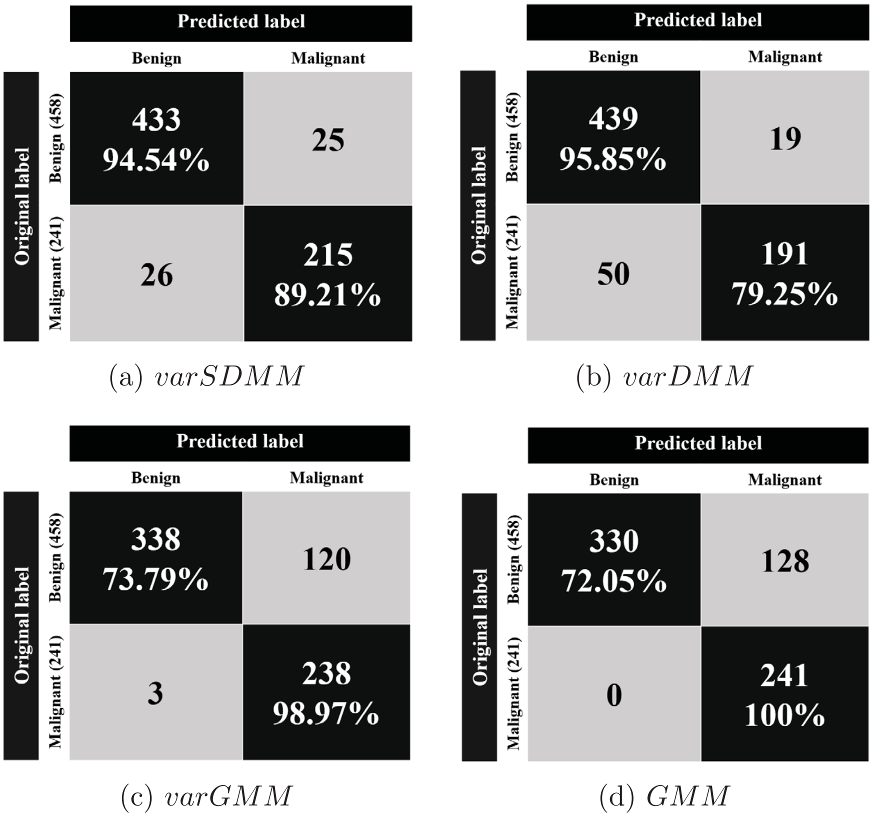


| Original label | Predicted label: Benign | Predicted label: Malignant |
|---|---|---|
| Benign (458) | 429<br>93.66% | 29 |
| Malignant (241) | 17 | 224<br>92.94% |

Figure 3.5: Wisconsin confusion matrix

Table 3.2: Model performance accuracy in Wisconsin dataset

| **Algorithm** | **varSSDMM** | **varSDMM** | **varDMM** | **varGMM** | **GMM** |
|---|---|---|---|---|---|
| **Accuracy(%)** | **93.4** | **92.7** | **90.1** | **82.7** | **81.6** |



| | | Predicted label | |
|---|---|---|---|
| | | Benign | Malignant |
| Original label | Benign (458) | 433<br>94.54% | 25 |
| | Malignant (241) | 26 | 215<br>89.21% |

(a) $varSDMM$

| | | Predicted label | |
|---|---|---|---|
| | | Benign | Malignant |
| Original label | Benign (458) | 439<br>95.85% | 19 |
| | Malignant (241) | 50 | 191<br>79.25% |

(b) $varDMM$

| | | Predicted label | |
|---|---|---|---|
| | | Benign | Malignant |
| Original label | Benign (458) | 338<br>73.79% | 120 |
| | Malignant (241) | 3 | 238<br>98.97% |

(c) $varGMM$

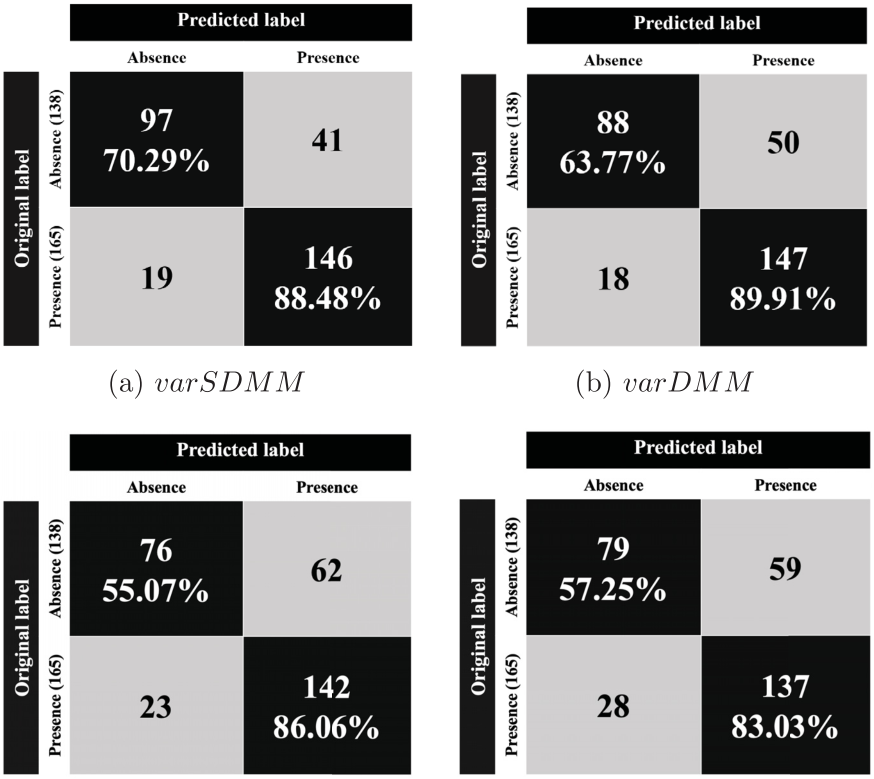
| | | Predicted label | |
|---|---|---|---|
| | | Benign | Malignant |
| Original label | Benign (458) | 330<br>72.05% | 128 |
| | Malignant (241) | 0 | 241<br>100% |

(d) $GMM$

Figure 3.6: Wisconsin confusion matrices for other algorithms

## 3.4 Cardiovascular Diseases (CVDs) Detection

Cardiovascular diseases (CVDs) as a wide assortment of disorders influencing the heart and veins, is perceived as the first reason for worldwide death. This leading explanation of mortality has ended the lives of 17.9 million individuals each year [56]. It is overwhelmingly costly to diagnose and control these illnesses due to the need for long-term treatment and pricey equipment. Thus, CVDs carries loads of expenses imposed to medicinal services and consequently government. However, considering the related risk factors of heart diseases such as obesity, tobacco use, low physical movement and diet, prevention could always be an essential approach. These days, complex data such as clinical history, biomarkers, pictures, signals and text are the source of analysis for doctors which could be a complicated task. Therefore, such diagnosis system could be error-prone, inaccurate and could put the patient in danger. In this situations, automation in clinical inference could be helpful [57]. In this part of our experiment, we evaluated our proposed model over a real and publicly available dataset [58] to predict heart disease existence based on specific characteristics of a person. There are 303 samples with 76 attributes being measured in this dataset, but all released experiments have utilized a subset of 14 features including age, sex, chest pain location, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results (normal, ST-T wave abnormality or left ventricular hypertrophy), maximum heart rate achieved, exercise induced angina, ST depression peak, the slope of the peak exercise ST segment (upsloping, flat or downsloping), number of major vessels coloured by fluoroscopy and type of defect (normal, fixed or reversible). The confusion matrix in Figure 3.7 confirms that most of the instances have been well classified, especially 92% of the heart disease presences has been detected which we can compare it with other algorithms' confusion matrices in Figure 3.8. The outcome of our assessment is demonstrated in Table 3.3 denoting the outperformance of varSSDMM with 82% overall accuracy. It is noteworthy to mention that reprocessing techniques such as min-max scaling and normalization has been performed on the dataset before applying our model.



| | | Predicted label | |
|---|---|---|---|
| | | Absence | Presence |
| Original label | Absence (138) | 97<br>70.29% | 41 |
| | Presence (165) | 19 | 146<br>88.48% |

(a) *varSDMM*

| | | Predicted label | |
|---|---|---|---|
| | | Absence | Presence |
| Original label | Absence (138) | 88<br>63.77% | 50 |
| | Presence (165) | 18 | 147<br>89.91% |

(b) *varDMM*

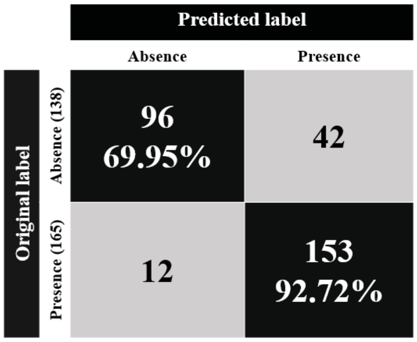
| | | Predicted label | |
|---|---|---|---|
| | | Absence | Presence |
| Original label | Absence (138) | 76<br>55.07% | 62 |
| | Presence (165) | 23 | 142<br>86.06% |

(c) *varGMM*

| | | Predicted label | |
|---|---|---|---|
| | | Absence | Presence |
| Original label | Absence (138) | 79<br>57.25% | 59 |
| | Presence (165) | 28 | 137<br>83.03% |

(d) *GMM*

Figure 3.8: Heart disease confusion matrices for other algorithms



| | | Predicted label | |
|---|---|---|---|
| | | Absence | Presence |
| Original label | Absence (138) | 96<br>69.95% | 42 |
| | Presence (165) | 12 | 153<br>92.72% |

Figure 3.7: Heart disease confusion matrix

Table 3.3: Model performance accuracy in heart dataset

| **Algorithm** | **varSSDMM** | **varSDMM** | **varDMM** | **varGMM** | **GMM** |
|---|---|---|---|---|---|
| **Accuracy(%)** | **82.2** | **80.1** | **77.5** | **71.9** | **71.3** |

## 3.5 Spam Detection

The forth real application we have tested our model on, is a text application. One of the serious research fields in information system security is spam detection. The concept of spam or unsolicited message is extended from product or website advertisements, money-making scams, pornography to chain letters. The most widely recognized form of spam is email spam which creates major problems such as lost productivity, financial damage and fraud. According to some references around 80% of emails are spam which brought about overwhelming financial losses of 50 billion dollars in 2005 [59]. Among all of the methodologies developed to stop spam, filtering is a significant and mainstream one. Applying machine learning and pattern recognition methods have significantly improved spam filtering compare to other user-defined rules [60], [61]. For our experiment, we obtained a challenging spam dataset from UCI machine learning repository, provided by Hewlett-Packard Labs [62]. The dataset has 4601 instances with 57 continuous input attributes plus the target column which denotes whether the e-mail was categorised spam (1) or not (0). Among all the instances, 39.4% of them (1813) are spam and 60.6% (2788) are non-spam. These attributes are obtained through a commonly used method called Bag of Words (BoW) [63] which is one of the effective data representation techniques in natural language processing. Majority of the attributes indicate the frequency of a particular word or character appearance in the email. In other words, each email is represented by its words ignoring grammar. 48 attributes contain the percentage of the respective word and 6 attributes include the percentage of the respective characters in the e-mail. The rest of them are the average length of continuous sequences of capital letters, the length of the longest continuous sequence of capital letters and the total number of capital letters in the e-mail. We carried out some pre-processing steps on the dataset before applying our model in order to enhance the final result. These steps include employing feature selection techniques which reduced the number of features to 45. Afterwards, min-max scaling and normalization is performed to further enhance our accuracy. In Figure 3.9, the outcome of our performance, represented in the confusion matrix, denoting better classification of most of the emails compare to other algorithms' confusion matrices in Figure 3.10. The summary of our model accuracy compared to four other algorithms represented in Table 3.4 that with 88.3% accuracy confirms the advantage of the model.

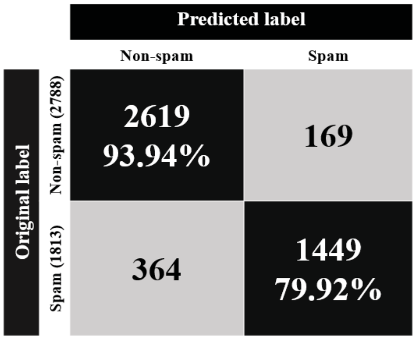
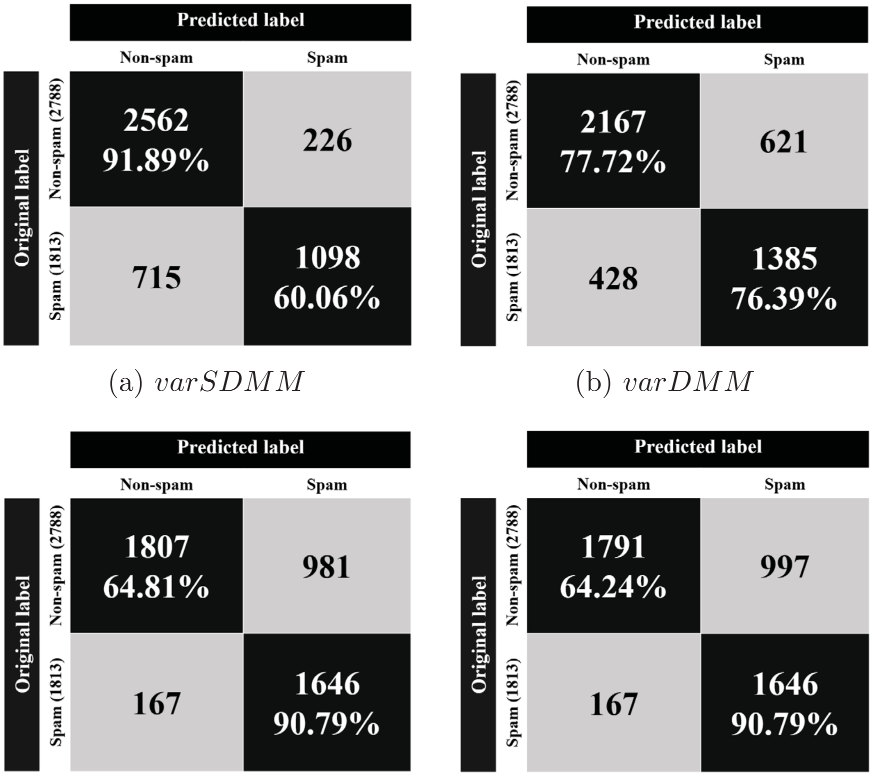


Figure 3.9: spam detection confusion matrix

Table 3.4: Model performance accuracy in spam detection dataset

| **Algorithm** | **varSSDMM** | **varSDMM** | **varDMM** | **varGMM** | **GMM** |
|---|---|---|---|---|---|
| **Accuracy(%)** | **88.3** | **79.5** | **77.2** | **75.0** | **74.7** |

Thus, we have evaluated our model on four different sizes of datasets and compared the results with three other variational models (varSDMM, varDMM, varGMM) as well as a deterministic model (GMM) to prove the potency and robustness of this model.



| Original label | Predicted label: Non-spam | Predicted label: Spam |
|---|---|---|
| Non-spam (2788) | 2562<br>91.89% | 226 |
| Spam (1813) | 715 | 1098<br>60.06% |

(a) *varSDMM*

| Original label | Predicted label: Non-spam | Predicted label: Spam |
|---|---|---|
| Non-spam (2788) | 2167<br>77.72% | 621 |
| Spam (1813) | 428 | 1385<br>76.39% |

(b) *varDMM*

| Original label | Predicted label: Non-spam | Predicted label: Spam |
|---|---|---|
| Non-spam (2788) | 1807<br>64.81% | 981 |
| Spam (1813) | 167 | 1646<br>90.79% |

(c) *varGMM*

| Original label | Predicted label: Non-spam | Predicted label: Spam |
|---|---|---|
| Non-spam (2788) | 1791<br>64.24% | 997 |
| Spam (1813) | 167 | 1646<br>90.79% |

(d) *GMM*

Figure 3.10: spam detection confusion matrices for other algorithms

Chapter 4

# Conclusion

The motivation behind this thesis was presenting the potential of clustering data using finite mixture models along with variational inference approach to further empowering the inference process. The significance of choosing a proper statistical model in order to describe data, was addressed. Although Gaussian mixture models are commonly used, they fail to model the data properly when dealing with non-Gaussian data like proportional ones. We suggested shifted-scaled Dirichlet mixture model, a generalized version of Dirichlet distribution family, as an excellent option for modeling asymmetric and proportional data such as normalized count vectors extracted from text and images. Also, we discussed the limitations of conventional estimation approaches such as Maximum Likelihood (ML) and Bayesian approach and what makes the variational Bayesian learning approach premier in data analysis tasks. Furthermore, we proposed a variational framework to learn the finite shifted-scaled Dirichlet mixture model. Using this model, parameters estimation was precisely accomplished avoiding the heavy cost of computation associated with conventional Bayesian strategies. Besides, the number of clusters was simultaneously detected as part of lower-bound maximization procedure. The experimental results have shown the validity of the proposed model, in terms of parameter approximation and detecting the true number of clusters, via several real world applications. The performance of our model in terms of accuracy is compared with four other algorithms, namely, variational learning of scaled Dirichlet mixture model (varSDMM), variational learning of Dirichlet mixture model (varDMM), variational learning of Gaussian mixture model (varGMM) and maximum likelihood learning of

Gaussian mixture model (GMM). Our model has proven to be more effective than other models in medical real applications including Malaria detection, breast cancer diagnosis and cardiovascular diseases as well as a challenging text application like spam detection. The proposed approach can be applied likewise to numerous different applications which contain proportional, asymmetric or scattered data such as text mining and natural language processing. In spite of the above mentioned advantages, we need to consider the limitations of the proposed model as well. Apart from complicated calculation of variational solution, it is very dependent to the initialization of hyperparameters and poor initialization values might considerably slow down the convergence speed. Therefore, we need to run the optimization several times with different initializations in order to detect a good maximum. Moreover, the proportional requirement of data in our model, makes it not applicable for some applications despite applying pre-processing techniques such as normalization. A future work could be devoted to make the approach less sensitive to the initialization of the hyperparameters via assuming a new level of prior distributions for the hyperparamters. Further enhancement of the model could include adding feature selection, upgrading to infinite model or developing online learning framework.

# Appendix A

# Proof of Equations (2.17), (2.18), (2.19) and (2.20)

In this section, we present the proof for (2.19), (2.20), (2.21) and (2.22): According to (2.18), we can rewrite the general expression of the variational solution $Q_s\big(\Theta_s\big)$ as below:

$$\ln Q_s\big(\Theta_s\big) = \big\langle \ln\ p\big(\mathcal{X}, \Theta\big)\big\rangle_{j \neq s} + const \tag{A.1}$$

In which those terms that are independent of the respective parameter in $Q_s\big(\Theta_s\big)$, are assimilated into the constant. Utilizing the equation (A.1) along with the logarithm of the joint distribution in (2.13), $p(\mathcal{X}, \Phi \mid \vec{\pi})$, we calculated variational solutions for each parameter as follow:

## A.1 Proof of (2.19) : Variational Solution to $Q\big(\mathcal{Z}\big)$

$$\begin{aligned}\ln Q(\mathcal{Z}_{ij}) = \mathcal{Z}_{ij}\Bigg\{ & \ln \pi_j + R_j - (D-1)\ln \overline{\tau}_j \\ & + \sum_{d=1}^{D}\Bigg[ -\frac{\overline{\alpha}_{jd}}{\overline{\tau}_j}\ln \overline{\beta}_{jd} + (\frac{\overline{\alpha}_{jd}}{\overline{\tau}_j} - 1)\ln x_{id}\Bigg] \\ & - (\alpha_j +)\ln\left(\sum_{d=1}^{D}(\frac{x_{id}}{\overline{\beta}_{jd}})^{\dfrac{1}{\overline{\tau}_j}}\right)\Bigg\} + const\end{aligned} \tag{A.2}$$

where $\quad R_j = \left\langle \ln \frac{\Gamma\big(\sum_{d=1}^{D}\alpha_{jd}\big)}{\prod_{d=1}^{D}\Gamma\big(\alpha_{jd}\big)} \right\rangle_{\alpha_{j1},\ldots,\alpha_{jD}} , \quad \overline{\alpha}_{jd} = \big\langle \alpha_{jd} \big\rangle = \frac{u^*_{jd}}{\nu^*_{jd}},$

However, $R_j$ is analytically intractable and we are not able to directly perform variational infernce. Thus, we need to approximate a lower bound for it that give us a closed-form expression. Obtaining a tractable approximation with applying second order Taylor series expansion in variational inference, has been effectively done in [64], [65]. Moreover, we can find the same function $R_j$ approximated using second order Taylor series in [45] that we will utilize it here. The approximation of $R_j$ around the expected values of $\vec{\alpha_j}$, represented by $\big(\overline{\alpha}_{j1},\ldots,\overline{\alpha}_{jD}\big)$ is defined as $\tilde{R}_j$ and denoted in (2.25). Now the equation in A.2 turns into a tractable expression after substituting $R_j$ by $\tilde{R}_j$ and we can obviously notice that the optimal solution estimation to $\mathcal{Z}$ takes the logarithmic form of 2.4 excluding the normalization constant. Therefore, we can rewrite $\ln Q\big(\mathcal{Z}\big)$ as follow:

$$\ln Q\big(\mathcal{Z}\big) = \sum_{i=1}^{N}\sum_{j=1}^{M}\mathcal{Z}_{ij}\ln \rho_{ij} + const \tag{A.3}$$

$$\begin{aligned}\ln \rho_{ij} &= \ln \pi_j + \tilde{R}_j - (D-1)\ln \overline{\tau}_j \\ &+ \sum_{d=1}^{D} \Bigg[ - \frac{\overline{\alpha}_{jd}}{\overline{\tau}_j} \ln \overline{\beta}_{jd} + (\frac{\overline{\alpha}_{jd}}{\overline{\tau}_j} - 1) \ln x_{id} \Bigg] \\ &- (\alpha_j +) \ln \left( \sum_{d=1}^{D} (\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}} \right)\end{aligned} \tag{A.4}$$

By taking the exponential of both sides in (A.3), we get:

$$Q\big(\mathcal{Z}\big) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \rho_{ij}^{Z_{ij}} \tag{A.5}$$

after applying normalization on the previous distribution, we obtain:

$$Q\big(\mathcal{Z}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{Z_{ij}}, \qquad r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M} \rho_{ij}} \tag{A.6}$$

note that the $\{r_{ij}\}$ are non-negative and sum to one. Thus, the standard solution for $Q(\mathcal{Z})$ can be derived as:

$$\big\langle Z_{ij} \big\rangle = r_{ij} \tag{A.7}$$

where $\{r_{ij}\}$ are equivalent to responsibilities in the conventional EM algorithms.

## A.2 Proof of (2.20) : Variational Solution to $Q\big(\vec{\alpha}\big)$

Since the parameters are considered statistically independent and there are M clusters in the mixture model, $Q(\vec{\alpha})$ can be factorized as follow:

$$Q(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{d=1}^{D} Q(\alpha_{jd}) \tag{A.8}$$

The logarithm of the optimized factor with respect to specific parameter $\alpha_{js}$ is calculated as follow:

$$\ln Q(\alpha_{js}) = \sum_{i=1}^{N} r_{ij}\mathcal{J}(\alpha_{js}) - \frac{\alpha_{js}}{\tau_j}\ln\beta_{js}\sum_{i=1}^{N} r_{ij} + \frac{\alpha_{js}}{\tau_j}\sum_{i=1}^{N} r_{ij}\ln x_{is} \\ - \alpha_{js}\sum_{i=1}^{N} r_{ij}\ln\left(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\right) + (u_{js}-1)\ln\alpha_{js} - \nu_{js}\alpha_{js} + const \tag{A.9}$$

where $\mathcal{J}(\alpha_{js}) = \left\langle \ln\frac{\Gamma(\alpha_s + \sum_{d\neq s}^{D}\alpha_{jd})}{\Gamma(\alpha_s)\prod_{d\neq s}^{D}\Gamma(\alpha_{jd})}\right\rangle_{\Theta\neq\alpha_{js}}$ is described as a function of $\alpha_{\alpha_{js}}$, which unfortunately doesn't have a closed-form solution. Therefore, same as $R_j$ in the section A, we need to approximate $\mathcal{J}(\alpha_{js})$ by finding a lower bound via Taylor series expansion about $\overline{\alpha}_{js}$ (the expected value of $\alpha_{js}$). The same function has been approximated in [45] (Appendix B) and we shall use the final result here:

$$\mathcal{J}(\alpha_{js}) \geq \overline{\alpha}_{js}\ln\alpha_{js}\Bigg\{\Psi\bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\bigg) - \Psi(\overline{\alpha}_{js}) + \sum_{d\neq s}^{D}\overline{\alpha}_{jd} \\ \times\Psi'\bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\bigg)(\langle\ln\alpha_{jd}\rangle - \ln\overline{\alpha}_{jd})\Bigg\} + const \tag{A.10}$$

after substituting this lower bound back into A.9, we get a new optimal solution to $\alpha_{js}$ as follow:

$$\ln Q(\alpha_{js}) = \sum_{i=1}^{N} r_{ij}\overline{\alpha}_{js}\ln\alpha_{js}\Bigg[\Psi\bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\bigg) - \Psi(\overline{\alpha}_{js}) \\ + \sum_{d\neq s}^{D}\Psi'\bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\bigg)\overline{\alpha}_{jd}(\langle\ln\alpha_{jd}\rangle - \ln\overline{\alpha}_{jd})\Bigg] \\ - \frac{\alpha_{js}}{\tau_j}\ln\beta_{js}\sum_{i=1}^{N} r_{ij} + \frac{\alpha_{js}}{\tau_j}\sum_{i=1}^{N} r_{ij}\ln x_{is} \\ - \alpha_{js}\sum_{i=1}^{N} r_{ij}\ln\left(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\right) + (u_{js}-1)\ln\alpha_{js} - \nu_{js}\alpha_{js} \\ = \ln\alpha_{js}(u_{js} + \varphi_{js} - 1) - \alpha_{js}(\nu_{js} + \vartheta_{js}) + const \tag{A.11}$$

where

$$\varphi_{jd} = \sum_{i=1}^{N} r_{ij}\overline{\alpha}_{jd}\Bigg[\psi\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) - \psi\big(\overline{\alpha}_{jd}\big) + \sum_{d\neq s}^{D}\psi'\Bigg(\sum_{d=1}^{D}\overline{\alpha}_{jd}\Bigg) \\ \times \overline{\alpha}_{js}\Big(\big\langle \ln\alpha_{js}\big\rangle - \ln\overline{\alpha}_{js}\Big)\Bigg] \tag{A.12}$$

$$\vartheta_{jd} = \sum_{i=1}^{N} r_{ij}\left[\frac{1}{\tau_j}\ln\frac{\beta_{js}}{x_{is}} + \ln\left(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\right)\right] \tag{A.13}$$

we can notice that (A.11) has gotten a logarithmic form of a Gamma function. if we take the exponential of both sides, we get:

$$Q(\alpha_{js}) \propto \alpha_{js}^{u_{js}+\varphi_{js}-1}e^{-\alpha_{js}(\nu_{js}+\vartheta_{js})} \tag{A.14}$$

Thus, we can derive the optimal solution to the hyperparamters $u_{js}$ and $\nu_{js}$ as:

$$u^{*}_{js} = u_{js} + \varphi_{js}, \quad \nu^{*}_{js} = \nu_{js} + \vartheta_{js}. \tag{A.15}$$

## A.3 Proof of (2.21) : Variational Solution to $Q\big(\vec{\beta}\big)$

Considering the assumption of parameter independence, for M cluster in the mixture model, $Q(\vec{\beta})$ can be factorized as follow:

$$Q(\vec{\beta}) = \prod_{j=1}^{M}\prod_{d=1}^{D} Q(\beta_{jd}) \tag{A.16}$$

The logarithm of the optimized factor with respect to specific parameter $\beta_{js}$ is calculated as follow:

$$\ln Q(\beta_{js}) = -\frac{\alpha_{js}}{\tau_j}\ln\beta_{js}\sum_{i=1}^{N} Z_{ij} - (\alpha_j+)\sum_{i=1}^{N} Z_{ij}\mathcal{F}(\beta_{js}) \\ + (h_{js}-1)\ln\beta_{js} + const \tag{A.17}$$

where

$$\mathcal{F}(\beta_{js}) = \left\langle \ln\Big(\sum_{d=1}^{D}\big(\frac{x_{id}}{\beta_{jd}}\big)^{\frac{1}{\tau_j}}\Big) \right\rangle \tag{A.18}$$

We can notice that $\mathcal{F}(\beta_{js})$ is analytically intractable and need to be approximated (see Appendix B). The lower bound is approximated about $\overline{\beta}_{js}$, as follow:

$$\mathcal{F}(\beta_{js}) \geq \big(\frac{x_{is}}{\overline{\beta}_{js}}\big)^{\frac{1}{\overline{\tau}_j}} \frac{-\ln \beta_{js}}{\overline{\tau}_j \sum_{d=1}^{D}\big(\frac{x_{id}}{\overline{\beta}_{jd}}\big)^{\frac{1}{\overline{\tau}_j}}} \tag{A.19}$$

By replacing this lower bound back into A.17, we have the following equation:

$$\begin{aligned}\ln(\beta_{js}) &= -\frac{\alpha_{js}}{\tau_j}\ln\beta_{js}\sum_{i=1}^{N} r_{ij} - (\alpha_j +)\sum_{i=1}^{N} r_{ij} \\ &\times \left[\big(\frac{x_{is}}{\overline{\beta}_{js}}\big)^{\frac{1}{\overline{\tau}_j}} \frac{-\ln \beta_{js}}{\overline{\tau}_j \sum_{d=1}^{D}\big(\frac{x_{id}}{\overline{\beta}_{jd}}\big)^{\frac{1}{\overline{\tau}_j}}}\right] + (h_{js}-1)\ln\beta_{js} + const \\ &= (h_{js}+\kappa_{js}-1)\ln\beta_{js} + const\end{aligned} \tag{A.20}$$

where

$$\kappa_{js} = \sum_{i=1}^{N} r_{ij}\left[\frac{-\bar{\alpha}_{js}}{\bar{\tau}_j} + \frac{\bar{\alpha_{js}}}{\bar{\tau}_j}\big(\frac{x_{is}}{\bar{\beta_{js}}}\big)^{\frac{1}{\bar{\tau}_j}} \frac{1}{\sum_{d=1}^{D}\big(\frac{x_{id}}{\bar{\beta_{jd}}}\big)^{\frac{1}{\bar{\tau}_j}}}\right] \tag{A.21}$$

We can notice that (A.20) has gotten a logarithmic form of a Beta distribution. By taking the exponential of both sides, we get:

$$Q(\beta_{js}) \propto \beta_{js}^{h_{js}+\kappa_{js}-1} \tag{A.22}$$

Therefore, we can extract the optimal solution to the hyperparamter $h_{js}$ as:

$$h^{*}_{js} = h_{js} + \kappa_{js} \tag{A.23}$$

## A.4 Proof of (2.22) : Variational Solution to $Q\big(\vec{\tau}\big)$

For M cluster in the mixture model, we can factorize $Q(\vec{\tau})$ as follow:

$$Q(\vec{\tau}) = \prod_{j=1}^{M} Q(\tau_j) \tag{A.24}$$

By taking logarithm of the optimized factor with respect to specific parameter $\tau_j$, we get:

$$\begin{aligned}\ln Q(\tau_j) = (1-D)\ln\tau_j \sum_{i=1}^{N} r_{ij} + \sum_{i=1}^{N} r_{ij} \sum_{d=1}^{D} \frac{\alpha_{jd}}{\tau_j}(\ln x_{id} - \ln\beta_{jd}) \\ - (\alpha_j +) \sum_{i=1}^{N} r_{ij}\mathcal{G}(\tau_j) + (q_j - 1)\ln\tau_j - s_j\tau_j + const\end{aligned} \tag{A.25}$$

where

$$\mathcal{G}(\tau_j) = \left\langle \ln\big(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\big) \right\rangle \tag{A.26}$$

Which is a function of $\tau_j$, again analytically intractable and need to be approximated (see Appendix B). We obtain the approximated lower bound, about $\overline{\tau}_j$, as follow:

$$\mathcal{G}(\tau_j) \geq \frac{-\ln\tau_j}{\overline{\tau}_j} \frac{\sum_{d=1}^{D}(\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}} \ln(\frac{x_{id}}{\overline{\beta}_{jd}})}{\sum_{d=1}^{D}(\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}}} + const \tag{A.27}$$

By substituting this lower bound back into A.25, we have the following equation:

$$\ln Q(\tau_j) = (1-D)\ln\tau_j \sum_{i=1}^{N} r_{ij} + \sum_{i=1}^{N} r_{ij} \sum_{d=1}^{D} \frac{\alpha_{jd}}{\tau_j} \tag{A.28}$$

$$\times (\ln x_{id} - \ln\beta_{jd}) + \frac{\ln\tau_j}{\overline{\tau}_j}(\alpha_j +) \sum_{i=1}^{N} r_{ij}$$

$$\times \frac{\sum_{d=1}^{D} (\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}} \ln(\frac{x_{id}}{\overline{\beta}_{jd}})}{\sum_{d=1}^{D} (\frac{x_{id}}{\overline{\beta}_{jd}})^{\frac{1}{\overline{\tau}_j}}} + (q_j - 1)\ln\tau_j - s_j\tau_j + const$$

$$= \ln\tau_j(q_j + \delta_j - 1) - \tau_j(s_j - \varrho_j)$$

where

$$\delta_j = \sum_{i=1}^{N} r_{ij} \left[ 1 - D + \frac{(\alpha_j +)}{\tau_j} \frac{\sum_{d=1}^{D} (\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}} \ln(\frac{x_{id}}{\beta_{jd}})}{\sum_{d=1}^{D} (\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}} \right] \tag{A.29}$$

$$\varrho_j = \sum_{i=1}^{N} r_{ij} \left[ \sum_{d=1}^{D} \frac{\alpha_{jd}}{{\tau_j}^2} \ln(\frac{x_{id}}{\beta_{jd}}) \right] \tag{A.30}$$

We can see that (A.28) has gotten a logarithmic form of a Gamma function. By taking the exponential of both sides, we get:

$$Q(\tau_j) \propto \tau_j^{q_j + \delta_j - 1} e^{-\tau_j(s_j - \varrho_j)} \tag{A.31}$$

So, we can obtain the optimal solution to the hyperparamters $q_j$ and $s_j$ as:

$$q_j^* = q_j + \delta_j, \quad s_j^* = s_j - \varrho_j. \tag{A.32}$$

Appendix **B**

# Proof of Equations (1.59) and (1.67)

## B.1 Proof of (1.59) : Lower Bond of $\mathcal{F}(\beta_{js})$

Let us define the function $\mathcal{F}(\beta_{js})$ as:

$$\mathcal{F}(\beta_{js}) = \ln\left( (\frac{x_{is}}{\beta_{js}})^{\frac{1}{\tau_j}} + \sum_{d\neq s}^{D} (\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}} \right) \tag{B.1}$$

Since $\mathcal{F}(\beta_{js})$ is a convex function with respect to $\ln \beta_{js}$, we can calculate its lower bound using first-order Taylor expansion of $\mathcal{F}(\beta_{js})$ for $\ln \beta_{js}$ at $\ln \beta_{js,0}$

as bellow:

$$\begin{aligned}\mathcal{F}(\beta_{js}) &\geq \mathcal{F}(\beta_{js,0}) + \frac{\partial \mathcal{F}(\beta_{js})}{\partial \ln \beta_{js}}|_{\beta_{js}=\beta_{js,0}}(\ln \beta_{js} - \ln \beta_{js,0}) \\ &= \ln\big((\frac{x_{is}}{\beta_{js,0}})^{\frac{1}{\tau_j}} + \sum_{d\neq s}^{D}(\frac{x_{id}}{\beta_{jd,0}})^{\frac{1}{\tau_j}}\big) \\ &+ \beta_{js,0}\frac{1}{(\frac{x_{is}}{\beta_{js,0}})^{\frac{1}{\tau_j}} + \sum_{d\neq s}^{D}(\frac{x_{id}}{\beta_{jd,0}})^{\frac{1}{\tau_j}}}(\frac{-1}{\tau_j}X_{is}^{\frac{1}{\tau_j}}\beta_{js,0}^{\frac{-1}{\tau_j}-1})(\ln \beta_{js} - \ln \beta_{js,0}) \\ &= (\frac{x_{is}}{\beta_{js,0}})^{\frac{1}{\tau_j}}\frac{-1}{\tau_j\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}} \times \ln \beta_{js}\end{aligned} \tag{B.2}$$

## B.2 Proof of (1.67) : Lower Bond of $\mathcal{G}(\tau_j)$

Let us define the function $\mathcal{G}(\tau_j)$ as:

$$\mathcal{G}(\tau_j) = \left\langle \ln\big(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_j}}\big) \right\rangle \tag{B.3}$$

Due to the convexity feature of function $\mathcal{G}(\tau_j)$ relative to $\ln \tau_j$, its lower bound can be calculated using first-order Taylor expansion of $\mathcal{G}(\tau_j)$ for $\ln \tau_j$

at $\ln \tau_{j,0}$ as follows:

$$\begin{aligned}
\mathcal{G}(\tau_j) &\geq \mathcal{G}(\tau_{j,0}) + \frac{\partial \mathcal{G}(\tau_j)}{\partial \ln \tau_j}|_{\tau_j=\tau_{j,0}}(\ln \tau_j - \ln \tau_{j,0}) \qquad \text{(B.4)} \\
&= \ln(\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_{j,0}}}) + \tau_{j,0}\frac{1}{\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_{j,0}}}} \\
&\times \frac{-1}{\tau_{j,0}^2}\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_{j,0}}}\ln(\frac{x_{id}}{\beta_{jd}})(\ln \tau_j - \ln \tau_{j,0}) \\
&= \frac{-\ln \tau_j}{\tau_{j,0}}\frac{\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_{j,0}}}\ln(\frac{x_{id}}{\beta_{jd}})}{\sum_{d=1}^{D}(\frac{x_{id}}{\beta_{jd}})^{\frac{1}{\tau_{j,0}}}} + const
\end{aligned}$$

# List of References

[1] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.

[2] Sharon X Lee, Geoffrey McLachlan, and Saumyadipta Pyne. Application of mixture models to large datasets. In *Big Data Analytics*, pages 57–74. Springer, 2016.

[3] Nizar Bouguila and Djemel Ziou. A probabilistic approach for shadows modeling and detection. In *IEEE International Conference on Image Processing 2005*, volume 1, pages I–329. IEEE, 2005.

[4] Ali Sefidpour and Nizar Bouguila. Spatial color image segmentation based on finite non-gaussian mixture models. *Expert Systems with Applications*, 39(10):8993–9001, 2012.

[5] Mohamad Mehdi, Nizar Bouguila, and Jamal Bentahar. Trustworthy web service selection using probabilistic models. In *2012 IEEE 19th International Conference on Web Services*, pages 17–24. IEEE, 2012.

[6] Nizar Bouguila and Djemel Ziou. Using unsupervised learning of a finite dirichlet mixture model to improve pattern recognition applications. *Pattern Recognition Letters*, 26(12):1916–1925, 2005.

[7] BS Everitt. An introduction to finite mixture distributions. *Statistical methods in medical research*, 5(2):107–127, 1996.

[8] Geoffrey J McLachlan and David Peel. *Finite mixture models.* John Wiley & Sons, 2004.

[9] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Novel mixtures based on the dirichlet distribution: application to data and image classification. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 172–181. Springer, 2003.

[10] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[11] Taoufik Bdiri and Nizar Bouguila. Positive vectors clustering using inverted dirichlet finite mixture models. *Expert Systems with Applications*, 39(2):1869–1882, 2012.

[12] Ram B Jain and Richard Y Wang. Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. *Analytical chemistry*, 80(12):4767–4772, 2008.

[13] Basim Alghabashi and Nizar Bouguila. Finite multi-dimensional generalized gamma mixture model learning based on mml. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1131–1138. IEEE, 2018.

[14] Dirk Husmeier. The bayesian evidence scheme for regularizing probability-density estimating neural networks. *Neural computation*, 12(11):2685–2717, 2000.

[15] Taoufik Bdiri, Nizar Bouguila, and Djemel Ziou. Variational bayesian inference for infinite generalized inverted dirichlet mixtures with feature selection and its application to clustering. *Applied Intelligence*, 44(3):507–525, 2016.

[16] Nizar Bouguila, Djemel Ziou, and Riad I Hammoud. A bayesian non-gaussian mixture analysis: Application to eye modeling. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[17] Dirk Husmeier, William D Penny, and Stephen J Roberts. An empirical evaluation of bayesian sampling with hybrid monte carlo for training neural network classifiers. *Neural Networks*, 12(4-5):677–705, 1999.

[18] Bjoern Bornkamp. Approximating probability densities by iterated laplace approximations. *Journal of Computational and Graphical Statistics*, 20(3):656–669, 2011.

[19] Lawrence J Brunner, Albert Y Lo, et al. Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics*, 17(4):1550–1566, 1989.

[20] Sami Bourouis, Mohamed Al Mashrgy, and Nizar Bouguila. Bayesian learning of finite generalized inverted dirichlet mixtures: Application to object classification and forgery detection. *Expert Systems with Applications*, 41(5):2329–2336, 2014.

[21] Wentao Fan and Nizar Bouguila. A variational component splitting approach for finite generalized dirichlet mixture models. In *2012 International Conference on Communications and Information Technology (ICCIT)*, pages 53–57. IEEE, 2012.

[22] Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning for finite dirichlet mixture models and applications. *IEEE transactions on neural networks and learning systems*, 23(5):762–774, 2012.

[23] Nizar Bouguila, Djemel Ziou, and Ernest Monga. Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, 16(2):215–225, 2006.

[24] Nizar Bouguila and Djemel Ziou. Unsupervised selection of a finite dirichlet mixture model: an mml-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):993–1009, 2006.

[25] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30, 1999.

[26] Hagai Attias. A variational bayesian framework for graphical models. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 209–215, 1999.

[27] Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and*

*Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

[28] Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

[29] Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning of finite dirichlet mixture models using component splitting. *Neurocomputing*, 129:3–16, 2014.

[30] Wentao Fan and Nizar Bouguila. Online variational learning of finite dirichlet mixture models. *Evolving Systems*, 3(3):153–165, 2012.

[31] Ines Channoufi, Sami Bourouis, Nizar Bouguila, and Kamel Hamrouni. Image and video denoising by combining unsupervised bounded generalized gaussian mixture modeling and spatial information. *Multimedia Tools and Applications*, 77(19):25591–25606, 2018.

[32] Sabri Boutemedjet, Djemel Ziou, and Nizar Bouguila. Model-based subspace clustering of non-gaussian data. *Neurocomputing*, 73(10-12):1730–1739, 2010.

[33] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.

[34] Nizar Bouguila and Walid ElGuebaly. A generative model for spatial color image databases categorization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 821–824. IEEE, 2008.

[35] Nizar Bouguila and Djemel Ziou. On fitting finite dirichlet mixture using ecm and mml. In *International conference on pattern recognition and image analysis*, pages 172–182. Springer, 2005.

[36] Nizar Bouguila and Djemel Ziou. Mml-based approach for finite dirichlet mixture estimation and selection. In *International workshop on machine learning and data mining in pattern recognition*, pages 42–51. Springer, 2005.

[37] Bromensele Samuel Oboh and Nizar Bouguila. Unsupervised learning of finite mixtures using scaled dirichlet distribution and its application to software modules categorization. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 1085–1090. IEEE, 2017.

[38] Hieu Nguyen, Muhammad Azam, and Nizar Bouguila. Data clustering using variational learning of finite scaled dirichlet mixture models. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pages 1391–1396. IEEE, 2019.

[39] Rua Alsuroji, Nizar Bouguila, and Nuha Zamzami. Predicting defect-prone software modules using shifted-scaled dirichlet distribution. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 15–18. IEEE, 2018.

[40] Juansnm José Egozcue and Vera Pawlowsky-Glahn. Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, 264(1):145–159, 2006.

[41] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons, 2011.

[42] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[43] Mahieddine M Ichir and Ali Mohammad-Djafari. A mean field approximation approach to blind source separation with l p priors. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.

[44] Giorgio Parisi. *Statistical field theory.* Addison-Wesley, 1988.

[45] Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning for finite dirichlet mixture models and applications. *IEEE transactions on neural networks and learning systems*, 23(5):762–774, 2012.

[46] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[47] David JC MacKay. Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469–505, 1995.

[48] World Health Organization et al. World malaria report. 2019.

[49] Malaria datasets — national library of medicine. https://lhncbc.nlm.nih.gov/publication/pub9932.

[50] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[51] Koffi Eddy Ihou and Nizar Bouguila. A new latent generalized dirichlet allocation model for image classification. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[52] Breast cancer statistics. https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics, Jul 2019.

[53] Fine needle aspiration (fna) biopsy of the breast. https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html.

[54] Uci machine learning repository, breast dataset. https://archive.ics.uci.edu/ml/datasets/breastcancerwisconsin(original).

[55] Ahmad Taher Azar and Shereen M El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8):2387–2403, 2013.

[56] Cardiovascular diseases (cvds). http://origin.who.int/cardiovascular_diseases/en/.

[57] Chayakrit Krittanawong, HongJu Zhang, Zhen Wang, Mehmet Aydar, and Takeshi Kitai. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, 69(21):2657–2664, 2017.

[58] heart disease data set. https://archive.ics.uci.edu/ml/datasets/HeartDisease.

[59] Enrico Blanzieri and Anton Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008.

[60] Levent Özgür and Tunga Güngör. Optimization of dependency and pruning usage in text classification. *Pattern analysis and applications*, 15(1):45–58, 2012.

[61] Ola Amayri and Nizar Bouguila. A study of spam filtering using support vector machines. *Artificial Intelligence Review*, 34(1):73–108, 2010.

[62] Uci machine learning repository: Spambase data set. `https://archive.ics.uci.edu/ml/datasets/spambase`.

[63] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[64] Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160–2173, 2011.

[65] Mark William Woolrich and Timothy E Behrens. Variational bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10):1380–1391, 2006.