Unsupervised Learning with Feature Selection Based on Multivariate McDonald's Beta Mixture Model for Medical Data Analysis

Darya Forouzanfar

Degree of Master of Applied Science
in the Department of
Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of
Master of Applied Science
"Information Systems Security" at
Concordia University
Montréal, Québec, Canada

June 2023

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared:

By: Darya Forouzanfar
Entitled: Unsupervised Learning with Feature Selection Based on Multivariate McDonald's Beta Mixture Model for Medical Data Analysis

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

Master of Applied Science **"Information Systems Security"**

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

Signed by the final examining committee:

Dr. Mohsen Ghafouri	_____	Examiner and Chair
Dr. Manar Amayri	_____	Examiner
Dr. Nizar Bouguila	_____	Supervisor

Approved by _____
Dr. Zachary Patterson, Graduate Program Director
Department of Concordia Institute for Information Systems Engineering (CIISE)

June 8, 2023 _____
Dr. Mourad Debbabi
Dean of Faculty of Engineering and Computer Science

Abstract

Unsupervised Learning with Feature Selection Based on Multivariate McDonald's Beta Mixture Model for Medical Data Analysis

Darya Forouzanfar
Concordia University, 2023

This thesis proposes innovative clustering approaches using finite and infinite mixture models to analyze medical data and human activity recognition. These models leverage the flexibility of a novel distribution, the multivariate McDonald's Beta distribution, offering superior capability to model data of varying shapes. We introduce a finite McDonald's Beta Mixture Model (McDBMM), demonstrating its superior performance in handling bounded and asymmetric data distributions compared to traditional Gaussian mixture models.

Further, we employ deterministic learning methods such as maximum likelihood via the expectation maximization approach and also a Bayesian framework, in which we integrate feature selection. This integration enhances the efficiency and accuracy of our models, offering a compelling solution for real-world applications where manual annotation of large data volumes is not feasible.

To address the prevalent challenge in clustering regarding the determination of mixture components number, we extend our finite mixture model to an infinite model. By adopting a nonparametric Bayesian technique, we can effectively capture the underlying data distribution with an unknown number of mixture components.

Across all stages, our models are evaluated on various medical applications, consistently demonstrating superior performance over traditional alternatives. The results of this research underline the potential of the McDonald's Beta distribution and the proposed mixture models in transforming medical data into actionable knowledge, aiding clinicians in making more precise decisions and improving health care industry.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Nizar Bouguila, for his unwavering support, guidance, and encouragement throughout my graduate studies. The opportunity to work under his supervision has been truly invaluable, and I have learned a great deal from his expertise and wisdom.

I am truly grateful for his dedication to my academic growth and for providing me with the necessary resources and knowledge to succeed in this journey.

Additionally, I would like to extend my heartfelt appreciation to Dr. Narges Manouchehri, for her continuous support, both academically and emotionally. Her insights, advice, and understanding have been instrumental in my progress, and I am grateful for her assistance and the positive impact she has had on my academic career.

Finally, I would like to thank my family and friends for their encouragement and support throughout this challenging yet rewarding experience. Their belief in me has been a source of strength and motivation, and I am grateful for the role they have played in my academic journey.

Contents

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**
1.1 Introduction and Related Work 1
1.2 Thesis Overview 3
1.3 Contributions 4
1.4 Publications 5

2 Finite Multivariate McDonald's Beta Mixture Model Learning Approach in Medical Applications **6**
2.1 Model Specification 6
2.1.1 Finite McDonald's Beta Distribution 6
2.1.2 Finite McDonald's Beta Mixture Model 8
2.2 Model Learning 8
2.2.1 Maximum Likelihood and Expectation Maximization . 8
2.2.2 Newton-Raphson Method 9
2.3 Model Complexity 13
2.3.1 Fisher Information for McDonald's Beta Mixture Model 13
2.3.2 Calculating Prior Distribution for MML Criterion . . . 14
2.4 Experimental Results 16
2.4.1 Targeting Treatment For Heart Disease Patients 18
2.4.2 Breast Tissue 19
2.4.3 Malaria Detection 22

3 A fully Bayesian Inference Approach for Multivariate McDonald's Beta Mixture Model with Feature Selection 24
3.1 Feature Saliency 24
3.1.1 Feature Selection in McDonald's Beta Mixture Model . 25
3.2 Model Learning 26
3.2.1 Bayesian Learning Framework 26
3.3 Experimental Results 32
3.3.1 Lung Cancer Analysis 32
3.3.2 Human Activity Recognition (HAR) 33

4 Bayesian Inference in Infinite Multivariate McDonald's Beta Mixture model 35
4.1 Model Learning 35
4.1.1 Bayesian Learning Framework 35
4.1.2 Extension to Infinite Mixture Model 36
4.2 Algorithm Overview 38
4.3 Experimental Results 39
4.3.1 Lung Cancer Analysis 39
4.3.2 Human Activity Recognition (HAR) 40

5 Conclusion 42

List of References 46

List of Figures

2.1 McDonald's Beta distribution 7
2.2 Message length plot for the Heart Disease dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 3. 18
2.3 Samples of benign and malignant breast tissue. Benign and malignant samples are represented in the first and second rows, respectively. 20
2.4 Message length plot for the Breast Tissue dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 2. 21
2.5 Samples of infected cells and uninfected cells. The samples are represented in the first and second rows. 21
2.6 Message length plot for the Malaria dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 2. 23

3.1 Lung Cancer cell image samples 33
3.2 Feature relevancy of eight random features in Lung cancer dataset across all components 34
3.3 Feature relevancy of eight random features in Human activity recognition dataset across all components 34

4.1 Sample of each type of lung cancer images 40

List of Tables

2.1 Results on Heart Disease dataset 17
2.2 Results on Breast Tissue Dataset 20
2.3 Results on Malaria Dataset 22

3.1 Results on Lung Cancer Dataset 32
3.2 Results on Human Activity Recognition Dataset 33

4.1 Results on Lung Cancer Dataset 40
4.2 Feature Relevancy Across All Components for Lung Cancer dataset 40
4.4 Feature Relevancy Across All Components for HAR dataset . 41
4.3 Results on HAR Dataset 41

List of Acronyms

BOVW: Bag of visual words

DP: Dirichlet processes

EM: Expectation-maximization

GMM: Gaussian mixture models

HAR: Human activity recognition

MCMC: Markov chain Monte Carlo

McDBD: McDonald's Beta distribution

FMcDBMM: Finite Multivariate McDonald's Beta mixture model

IMcDBMM: InFinite Multivariate McDonald's Beta mixture model

ML: Maximum likelihood

MML: Minimum message length

SIFT: Scale-invariant feature transform

FN: False negatives

TN: True negatives

FP: False positives

TP: True positives

WHO: World Health Organization

Chapter 1

Introduction

1.1 Introduction and Related Work

Machine learning and data mining strategies have increasingly garnered interest for their remarkable capability in modelling and deciphering data from a diverse range of fields, such as pattern recognition, computer vision and image processing [1, 2]. As an unsupervised learning method, clustering categorizes data into different groups with similar characteristics. The main idea of clustering is to group unlabelled data so that data points within a cluster have more similarities than those in other clusters [3, 4].

Within statistical learning techniques, finite mixture models have showcased their effectiveness in modelling complex data sets by theorizing that each observation originates from one of several distinct groups or components [5–9]. However, selecting the most proper probability distribution is required to characterize the components adequately. Gaussian Mixture Models (GMMs) have been popularly utilized for clustering tasks and have demonstrated remarkable fitting capabilities in various applications [10]. However, GMMs are not the best choice in the presence of non-Gaussian data and asymmetrical structures [11–15]. In light of this, alternative distributions such as Beta-Liouville [16–19], Dirichlet [20–22] and generalized Dirichlet [23–25] have been explored, proving more suitable for data clustering.

This thesis proposes a novel finite mixture model and develops it based on an extended version of Beta distribution called McDonald's Beta distribu-

tion [26]. Our motivation is its flexibility, as this distribution has four shape parameters that provide good potential to fit asymmetric and non-Gaussian data that, with promising results on real-world datasets, can be considered an alternative to Gaussian distribution.
Implementing mixture models involves two challenging aspects: estimating model parameters and determining the model's complexity avoiding underfitting or overfitting [5]. Various deterministic and Bayesian approaches have been developed to handle the former challenges. Deterministic approaches, such as maximum likelihood (ML) estimation via the expectation-maximization (EM) algorithm, are prized for their simplicity and low computational complexity. However, the EM algorithm does have its drawbacks, including convergence to a local maximum, dependency on initialization, and overfitting problems [27]. On the other hand, Bayesian inference, powered by advancements in computational methods, offers a potential alternative that can provide more accurate results [28].

The main concept of the Bayesian method is to extract properties of the probability distribution from data using Bayes' theorem, updating the prior beliefs about parameters based on insights drawn from the observations to determine the posterior [29, 30]. This approach, which relies heavily on sampling techniques and employs Markov Chain Monte Carlo (MCMC) for Bayesian inference [31], is the foundation for the framework introduced in the third chapter of this thesis. Moreover, we utilize Gibbs sampling within the Metropolis-Hastings algorithm for estimating the parameters of the finite multivariate McDonald's Beta mixture model. Furthermore, to address the second major challenge in implementing mixture models - accurately determining the number of clusters - we extend our finite model to an infinite mixture model using a mixture of Dirichlet processes [32–34] for a non-parametric Bayesian framework which performs simultaneous parameter estimation and model selection.

In addition, in data mining and machine learning, data clustering presents multiple challenges when dealing with high-dimensional data due to the issue of data sparsity and the presence of irrelevant features. As such, feature selection is critical in enhancing clustering performance in those cases [35, 36]. The primary aim of feature selection is to identify and diminish the impact of irrelevant features, which do not add significant information to the actual cluster structure [37].

However, executing the automatic selection of relevant features within unsupervised learning scenarios is challenging. This complexity arises as inferences must be made simultaneously regarding the relevant features and the clustering structure [38, 39]. The fundamental notion suggested in [35] was that each feature is derived from a mixture of two univariate distributions. The first distribution is presumed to generate relevant features and varies for each cluster. Also, the second distribution, common to all clusters and independent of class labels, is postulated to generate irrelevant features. Therefore, we have integrated an unsupervised feature selection approach in the proposed mixture models explained in chapters 3 and 4 to address high-dimensional data challenges and improve our model performance.

1.2 Thesis Overview

This thesis is organized as follows:

- Chapter 2: We propose a novel finite mixture model based on McDonald's Beta distribution to handle asymmetric and non-Gaussian data. We implement ML via EM to learn our model and introduce the Minimum Message Length criterion for model [40, 41]. We validated our model on three challenging medical applications.

- Chapter 3: We tackle the limitations of deterministic methods by employing Bayesian inference for model parameter estimation. We utilize the Markov Chain Monte Carlo technique, which includes Gibbs sampling and the Metropolis-Hastings method, within our FMcDBMM. In addition, we integrate feature selection to determine feature saliency and evaluated our model on real world datasets in medical applications for lung cancer and human activity recognition (HAR).

- Chapter 4: We introduce an extension to our model, the infinite multivariate McDonald's Beta mixture model. This model employs a mixture of DP to automatically estimate the complexity of the model and determine the number of components. We applied the model on the same datasets as chapter 3 to compare the results and demonstrate the enhancement.

- Chapter 5: We bring together our key findings and summarize our

contributions. We also take the opportunity to underscore some of the challenges encountered during our research.

1.3 Contributions

In this thesis, our main contributions can be summarized as follows:

- Propose a novel finite and infinite mixture model. This work includes the introduction of a new finite mixture model based on the extended version of the Beta distribution, McDonald's Beta distribution. This approach offers a more flexible fitting for non-Gaussian and asymmetric data.

- Focus on deterministic learning methods for finite mixture models such as ML via EM using Newton Raphson's method, as well as non-deterministic methods such as Bayesian inference approaches. We particularly focus on the use of Markov Chain Monte Carlo (MCMC) techniques, including Gibbs sampling and Metropolis-Hastings (M-H) methods.

- Introduce an extension to the finite mixture model, the Infinite Multivariate McDonald's Beta Mixture Model (IMcDBMM). This model utilizes a mixture of Dirichlet processes, enabling automatic estimation of model's complexity in clustering tasks.

- Provide a rigorous evaluation of our proposed models using real-world data sets from medical applications and human activity recognition. This includes a comparative analysis of our models' performance against similar alternatives, demonstrating their effectiveness in diverse practical scenarios.

- Utilize advanced feature extraction techniques such as SIFT and BOVW in our applications, further enhancing our models' ability to handle complex data.

1.4 Publications

This thesis consists of three manuscripts accepted as conference papers. We hereby list them:

- Chapter 2: Darya Forouzanfar, Narges Manouchehri, Nizar Bouguila, Finite Multivariate McDonald's Beta Mixture Model Learning Approach in Medical Applications, in proceedings of "The 38th ACM/SIGAPP Symposium on Applied Computing" [42].

- Chapter 3: Darya Forouzanfar, Narges Manouchehri, Nizar Bouguila, A fully Bayesian Inference Approach for Multivariate McDonald's Beta Mixture Model with Feature Selection, in proceedings of "The 9th International Conference on Control, Decision and Information Technologies CoDIT 2023" [43].

- Chapter 4: Darya Forouzanfar, Narges Manouchehri, Nizar Bouguila, Bayesian Inference in Infinite Multivariate McDonald's Beta Mixture Model, in proceedings of "The 22nd International Conference on Artificial Intelligence and Soft Computing ICAISC 2023" [44].

Chapter 2

Finite Multivariate McDonald's Beta Mixture Model Learning Approach in Medical Applications

In this chapter, we first introduce a finite mixture model with a generalization of Beta distribution called the McDonald's Beta Mixture Model (McDBMM). Subsequently, we delve into the estimation of the McDBMM parameters using the Maximum Likelihood (ML) and Expectation-Maximization (EM) algorithms. The Newton-Raphson method, serving as an iterative approach, assists us in the computation of the updated parameters. Moreover, we also present the Minimum Message Length (MML) as the model complexity approach used in our work.

2.1 Model Specification

2.1.1 Finite McDonald's Beta Distribution

McDonald's Beta distribution (McDBD) is a generalized version of Beta distribution and has four shape parameters [26]. To describe it, let's assume a D-dimensional data point $\vec{X}_n = \big(x_{n1}, \ldots, x_{nd}\big)$ following McDBD where $0 \leq x_{nd} \leq q_{jd}$, $q_{jd} > 0$ and $d = 1, \ldots, D$. The four shape parameters of McDBD are as follows: $\vec{a}_j = \big(a_{j1}, \ldots, a_{jd}\big)$, $\vec{b}_j = \big(b_{j1}, \ldots, b_{jd}\big)$, $\vec{p}_j = \big(p_{j1}, \ldots, p_{jd}\big)$, $\vec{q}_j = \big(q_{j1}, \ldots, q_{jd}\big)$ such that $a_{jd} > 0$, $b_{jd} > 0$, $p_{jd} > 0$ for $d = 1, \ldots, D$. So,

we express the joint density function of this observation as follows:

$$p(\vec{X_n} \mid \vec{a_j}, \vec{b_j}, \vec{p_j}, \vec{q_j}) = \prod_{d=1}^{D} \frac{p_{jd} {x_{nd}}^{a_{jd}p_{jd}-1} (q_{jd}{}^{p_{jd}} - {x_{nd}}^{p_{jd}})^{b_{jd}-1}}{{q_{jd}}^{p_{jd}(a_{jd}+b_{jd}-1)} B(a_{jd}, b_{jd})} \tag{2.1}$$

where:

$$B(a_{jd}, b_{jd}) = \int_0^1 t^{a_{jd}-1}(1-t)^{b_{jd}-1} dt = \frac{\Gamma(a_{jd})\Gamma(b_{jd})}{\Gamma(a_{jd}+b_{jd})} \tag{2.2}$$

In this chapter, we assume that $q = 1$ to make the support between zero and one and we have:

$$p(\vec{X_n} \mid \vec{a_j}, \vec{b_j}, \vec{p_j}, \vec{q_j}) = \prod_{d=1}^{D} \frac{p_{jd} {x_{nd}}^{a_{jd}p_{jd}-1} (1 - {x_{nd}}^{p_{jd}})^{b_{jd}-1}}{B(a_{jd}, b_{jd})} \tag{2.3}$$

In fact, if we set q and p of McDBD equal to one, we will obtain Beta distribution. We demonstrate some examples of this distribution in Fig. 2.1.

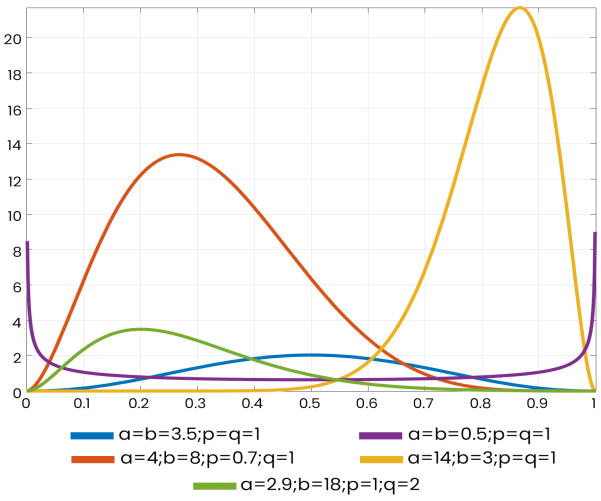

Figure 2.1: McDonald's Beta distribution

2.1.2 Finite McDonald's Beta Mixture Model

To formulate a finite McDonald's Beta mixture model for $\vec{X}_n$ assuming that there are M components, we have:

$$p(\vec{X}_n \mid \Theta) = \sum_{j=1}^{M} w_j p\big(\vec{X}_n \mid \vec{\theta}_j\big) \tag{2.4}$$

w_j and $\vec{\theta}_j = (\vec{a}_j, \vec{b}_j, \vec{p}_j)$ are respective weight and set of parameters of component j, where $j = 1, \ldots, M$. $\Theta = \{\vec{w}, \vec{\theta}\}$ is the complete set of mixture parameters where $\vec{w} = (w_1, \ldots, w_M)$, $\sum_{j=1}^{M} w_j = 1$ and $w_j >= 0$ for $j = 1, \ldots, M$ and $\vec{\theta} = (\vec{\theta}_1, \ldots, \vec{\theta}_M)$.
To model $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ as a dataset with N D-dimensional independent and identically distributed observations, we have:

$$p(\mathcal{X} \mid \Theta) = \prod_{n=1}^{N} \left[\sum_{j=1}^{M} w_j p\big(\vec{X}_n \mid \vec{\theta}_j\big) \right] \tag{2.5}$$

$$= \prod_{n=1}^{N} \left[\sum_{j=1}^{M} w_j \prod_{d=1}^{D} \frac{p_{jd} {x_{nd}}^{a_{jd}p_{jd}-1} (1 - {x_{nd}}^{p_{jd}})^{b_{jd}-1}}{B(a_{jd}, b_{jd})} \right]$$

2.2 Model Learning

2.2.1 Maximum Likelihood and Expectation Maximization

To tackle the model estimation problem, the parameters which maximize the probability density function of data are determined using ML and EM frameworks. EM is an approach to obtain maximum likelihood estimation in the case of having latent variables. EM has two main steps, The first step is to estimate the values for the latent variables, and the second is to optimize the model using ML algorithm. ML is an estimation procedure to find the mixture model parameters that maximize log-likelihood function which is defined by:

$$L(\Theta, \mathcal{X}) = \log p(\mathcal{X} \mid \Theta) = \sum_{n=1}^{N} \log \sum_{j=1}^{M} w_j p(\vec{X}_n \mid \vec{\theta}_j) \tag{2.6}$$

Each $\vec{X}_n$ belongs to one of the j components hence, we propose a vector $\vec{Z}_n = (Z_{n1}, ..., Z_{nj})$ such that $Z_{nj} = 1$ for $j = (1, \ldots, M)$ if $\vec{X}_n$ belongs to component j, else 0 and $\sum_{j=1}^{M} Z_{nj} = 1$. For $\mathcal{X}$, we define a set of membership vectors $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$. Also, we should mention that we assign each vector $\vec{X}_n$ to one of the M clusters by its posterior probability given by:

$$\hat{Z}_{nj} = p\big(j \mid \vec{X}_n, \vec{\theta}_j\big) = \frac{w_j p(\vec{X}_n, \vec{\theta}_j)}{\sum_{j=1}^{M} w_j p(\vec{X}_n, \vec{\theta}_j)} \tag{2.7}$$

Therefore, the complete log-likelihood is given as follows:

$$\begin{aligned} L(\Theta, \mathcal{Z}, \mathcal{X}) &= \sum_{n=1}^{N}\sum_{j=1}^{M} \hat{Z}_{nj}(\log w_j + \log p\big(\vec{X}_n \mid \vec{\theta}_j\big)) \qquad (2.8)\\ &= \sum_{n=1}^{N}\sum_{j=1}^{M} \hat{Z}_{nj}(\log w_j + \sum_{d=1}^{D}[\log p_{jd} + \\ &(a_{jd}p_{jd} - 1)\log x_{nd} + (b_{jd} - 1)\log(1 - x_{nd}^{p_{jd}}) + \\ &\log\Gamma(a_{jd} + b_{jd}) - \log\Gamma(a_{jd}) - \log\Gamma(b_{jd})] \end{aligned}$$

In the next step, we are going to maximize the complete log-likelihood which is computed in (2.8) by calculating the gradient of the log-likelihood with respect to parameters:

$$\frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \Theta} = 0 \tag{2.9}$$

2.2.2 Newton-Raphson Method

As (2.9) doesn't have a closed-form solution, we use Newton-Raphson as an iterative approach to update parameters. The Newton-Raphson method is an effective technique for estimating a function by making a local quadratic approximation based on information from the current point and then jumping to the minimum of that approximation. G as the gradients is the first derivative of $L(\Theta, \mathcal{Z}, \mathcal{X})$ with respect to the parameters. H as Hessian matrix is the second and mixed derivatives of $L(\Theta, \mathcal{Z}, \mathcal{X})$ with respect to the parameters. So, we update parameters as follows:

$$\hat{a_j}^{new} = \hat{a_j}^{old} - {H_j}^{-1} G_j \tag{2.10}$$
$$\hat{b_j}^{new} = \hat{b_j}^{old} - {H_j}^{-1} G_j$$
$$\hat{p_j}^{new} = \hat{p_j}^{old} - {H_j}^{-1} G_j$$

By calculating derivatives with respect to a_{jd}, b_{jd} and p_{jd} where $\psi(X) = \frac{\Gamma'(X)}{\Gamma(X)}$, we have:

$$G(a) = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd}} = \tag{2.11}$$
$$\sum_{n=1}^{N} \hat{Z}_{nj} [p_{jd} \log x_{nd} + \Psi(a_{jd} + b_{jd}) - \Psi(a_{jd})]$$
$$G(b) = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd}} \tag{2.12}$$
$$= \sum_{n=1}^{N} \hat{Z}_{nj} [\log(1 - x_{nd}^{p_{jd}}) + \Psi(a_{jd} + b_{jd}) - \Psi(b_{jd})]$$
$$G(p) = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd}} \tag{2.13}$$
$$= \sum_{n=1}^{N} \hat{Z}_{nj} [\frac{1}{p_{jd}} + a_{jd} \log(x_{nd}) + \frac{(1 - b_{jd}) \log(x_{nd}) x_{nd}^{p_{jd}}}{1 - x_{nd}^{p_{jd}}}]$$

To calculate Hessian matrix, we compute second and mixed derivatives of log-likelihood function.

- Derivatives with respect to a_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd}^2} = \sum_{n=1}^{N} \hat{Z}_{nj} [\psi'(a_{jd} + b_{jd}) - \psi'(a_{jd})] \tag{2.14}$$
$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{2.15}$$

- Derivatives with respect to a_{jd}, b_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd} \partial b_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\psi'(a_{jd} + b_{jd})] \tag{2.16}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_s} \partial b_{jd_t}} = 0, d_s \neq d_t \tag{2.17}$$

- Derivatives with respect to a_{jd}, p_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd} \partial p_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\log(x_{nd})] \tag{2.18}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_s} \partial p_{jd_t}} = 0, d_s \neq d_t \tag{2.19}$$

- Derivatives with respect to b_{jd}, a_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd} \partial a_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\psi'(a_{jd} + b_{jd})] \tag{2.20}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{2.21}$$

-Derivatives with respect to b_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd}^2} = \sum_{n=1}^{N} \hat{Z}_{nj}[\psi'(a_{jd} + b_{jd}) - \psi'(b_{jd})] \tag{2.22}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_s} \partial b_{jd_t}} = 0 \tag{2.23}$$

- Derivatives with respect to b_{jd}, p_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd} \partial p_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\frac{\log(x_{nd}) x_{nd}^{p_{jd}}}{x_{nd}^{p_{jd}} - 1}] \tag{2.24}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_s} \partial p_{jd_t}} = 0, d_s \neq d_t \tag{2.25}$$

- Derivatives with respect to p_{jd}, a_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd} \partial a_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\log(x_{nd})] \tag{2.26}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{2.27}$$

- Derivatives with respect to p_{jd}, b_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd} \partial b_{jd}} = \sum_{n=1}^{N} \hat{Z}_{nj}[\frac{\log(x_{nd}) x_{nd}^{p_{jd}}}{x_{nd}^{p_{jd}} - 1}] \tag{2.28}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd_s} \partial b_{jd_t}} = 0, d_s \neq d_t \tag{2.29}$$

- Derivatives with respect to p_{jd}:

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd}^2} = \sum_{n=1}^{N} \hat{Z}_{nj}[\frac{(1 - b_{jd}) x_{nd}^{p_{jd}} \{\log(x_{nd})\}^2}{(1 - x_{nd}^{p_{jd}})^2} - \frac{1}{p_{jd}^2}] \tag{2.30}$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial p_{jd_s} \partial p_{jd_t}} = 0, d_s \neq d_t \tag{2.31}$$

Our Hessian matrix is a $3D$ by $3D$ matrix as shown below:

$$H_j = \begin{bmatrix} H_{(a_{jd}, a_{jd})} & H_{(a_{jd}, b_{jd})} & H_{(a_{jd}, p_{jd})} \\ H_{(b_{jd}, a_{jd})} & H_{(b_{jd}, b_{jd})} & H_{(b_{jd}, p_{jd})} \\ H_{(p_{jd}, a_{jd})} & H_{(p_{jd}, b_{jd})} & H_{(p_{jd}, p_{jd})} \end{bmatrix} \tag{2.32}$$

To estimate the values of mixing proportion we will follow this equation:

$$w_j = \frac{\sum_{n=1}^{N} p\big(j \mid \vec{X}_n, \vec{\theta}_j\big)}{N} \tag{2.33}$$

In order to have an optimal performance of our model, initialization should be done adequately to avoid convergence to a local maximum which cannot be guaranteed using the EM method. We use K-means algorithm to initialize mixing proportions.

2.3 Model Complexity

Model selection helps to obtain the optimal number of clusters that best describes the data. Also, We need to identify the number of mixture components in the model to implement the EM algorithm. In this section, we will use a model selection technique called the minimum message length (MML). Deterministic model selection techniques are based on Bayesian method or information theory concepts and MML as a model selection approach is taking advantage of both [45]. According to that, we decided to choose MML as model selection approach for this work. Regarding information theory, the optimal number of clusters requires minimum information to transmit the data from sender to receiver efficiently. MML is based on that idea, and for a mixture of distributions, it is defined below as:

$$\text{MML} = -\log(\frac{h(\Theta)p(\mathcal{X} \mid \Theta)}{\sqrt{\mid F(\Theta) \mid}}) + N_p(-\frac{1}{2}\log(12) + \frac{1}{2}) \tag{2.34}$$

$h(\Theta)$ is prior probability distribution. N_p is the number of free parameters and equal to $(M(2D+1)) - 1$. $p(\mathcal{X} \mid \Theta)$ is the complete data log-likelihood and $|F(\Theta)|$ is the determinant of the Fisher information matrix which is defined by taking the second derivative of the negative log-likelihood.

2.3.1 Fisher Information for McDonald's Beta Mixture Model

Fisher matrix also named the curvature matrix, explains the curvature of the likelihood function around its maximum and is defined as the expected value of the negative of the Hessian matrix. In summary, it is the expected value of the negative of the second derivative of the log-likelihood function. Considering a mixture model, according to [46], Fisher information matrix can be calculated after the assignment of data vectors to their specific clusters. The determinant of the Fisher information matrix is defined below as [47]:

$$|F(\Theta)| = |F(\vec{w})| \prod_{j=1}^{M} |F(\vec{\theta_j})| \tag{2.35}$$

$|F(\vec{\theta_j})|$ is the determinant of Fisher information of $\vec{\theta}_j = (\vec{a}_j, \vec{b}_j, \vec{p}_j)$ and $|F(\vec{w})|$ is the determinant of Fisher information of mixing parameters w_j.

Therefore, considering generalized Bernoulli process with a series of trials where each has M possible results for M clusters, determinant of the Fisher information matrix can be computed as below:

$$|F(\vec{w})| = \frac{N^{M-1}}{\prod_{j=1}^{M} w_j} \tag{2.36}$$

Fisher information for our mixture will be:

$$\log(|F(\Theta)|) = \tag{2.37}$$
$$(M-1)\log(N) - \sum_{j=1}^{M} \log(w_j) + \sum_{j=1}^{M} \log(|F(\vec{\theta_j})|)$$

2.3.2 Calculating Prior Distribution for MML Criterion

For calculating MML criterion, we need to calculate prior distribution $h(\Theta)$ for model's parameters. Since these parameters are independent, we define $h(\Theta)$ as follow [48]:

$$h(\Theta) = h(\vec{w})h(\vec{a})h(\vec{b})h(\vec{p}) \tag{2.38}$$

Regarding the suitability of mixing parameters in modelling proportional vectors and the fact that $\sum_{j=1}^{M} w_j = 1$, we consider probability density of $h(\vec{w})$ to follow a Dirichlet distribution where $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_M)$:

$$h(w_1, w_2, \ldots, w_M) = \frac{\Gamma(\sum_{j=1}^{M} \eta_j)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} w_j^{\eta_j - 1} \tag{2.39}$$

A uniform prior for the parameter $\eta, (\eta_1 = 1, ..., \eta_M = 1)$ allows us to simplify (2.39), and we can calculate it as follows:

$$h(\vec{w}) = (M-1)! \tag{2.40}$$

For calculating $h(a)$, we suppose that dimensions are independent, so we have:

$$h(\vec{a}) = \prod_{j=1}^{M} \prod_{d=1}^{D} h(a_{jd}) \tag{2.41}$$

We assume that we do not have prior knowledge about parameter a_{jd}. So, it should have minimal effect on the posterior and we use following simple uniform prior which proved to have good results [49, 50]. The same process will be done for calculating $h(b_{jd})$ and $h(p_{jd})$, and:

$$h(a_{jd}) = e^{-6}\frac{a_{jd}}{||a_j||}, h(b_{jd}) = e^{-6}\frac{b_{jd}}{||b_j||}, h(p_{jd}) = e^{-6}\frac{p_{jd}}{||p_j||} \tag{2.42}$$

Log of prior is given by:

$$\begin{aligned}\log(h(\Theta)) = &-D\sum_{j=1}^{M}\log(||a_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(a_{jd}) \\ &- D\sum_{j=1}^{M}\log(||b_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(b_{jd}) \\ &- D\sum_{j=1}^{M}\log(||p_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(p_{jd}) \\ &+ \sum_{j=1}^{M-1}\log(j) - 18MD\end{aligned} \tag{2.43}$$

Algorithm 1 Full Learning Algorithm

1. Input X and the number of clusters M.
2. Use K-Means algorithm to initialize the M clusters.
3. Initialize the parameters.
   Repeat
4. EM algorithm
   (a) E step: Compute $\hat{Z}_{nj}$
   (b) M step: Update the parameters and weight.
   (c) If $w_j < \epsilon$ then delete component j return to E step.

   until
   Convergence

5. MML
   (a) Calculate the criterion of MML(M).
   (b) Find the optimal M^* i.e. $M^* = argmin_M$ MML(M).

2.4 Experimental Results

In this section, we test the performance of FMcDBMM and compare it with Gaussian finite mixture models. To validate the robustness of our proposed model, we apply it to three real medical applications: Targeting treatment for heart disease patients, breast tissue analysis and malaria detection. As one of our distribution assumptions is that all input values are in the range of (0,1), we normalize our datasets using the min-max method as below:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.44}$$

Also, we consider four metrics of accuracy, precision, recall and F1-score to evaluate model's robustness, where TP, TN, FP and FN are the respective

number of true positives, true negatives, false positives, and false negatives:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F1 - score &= 2 * \frac{Precision * Recall}{Precision + Recall}
\end{aligned}
\tag{2.45}
$$

Table 2.1: Results on Heart Disease dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
McDBMM	72.05	58.08	58.08	58.08
GMM	64.79	47.19	47.19	47.19

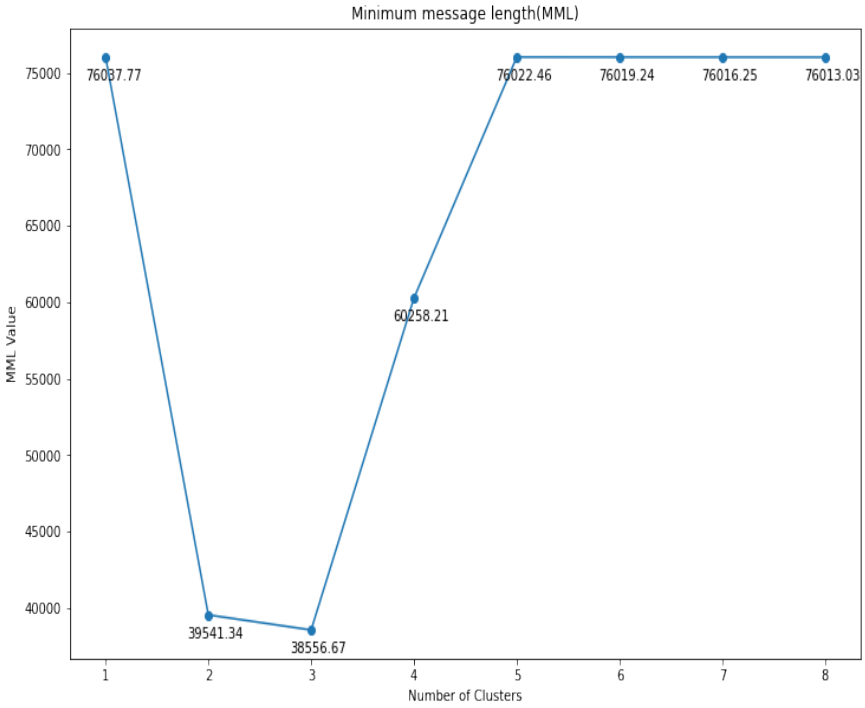

Figure 2.2: Message length plot for the Heart Disease dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 3.

2.4.1 Targeting Treatment For Heart Disease Patients

In this part of our experiment, we are going to cluster anonymized data of patients who have been diagnosed with heart disease. We analyze data from V.A. Medical Center in Long Beach, CA [51]. This publicly available 3-components dataset contains 303 instances and nine attributes plus the target. The attributes describing the dataset are age of the patient, gender, chest pain type, resting blood pressure (in mm Hg on admission to the hospital), Serum cholesterol in mg/dl, fasting blood sugar $>$ 120 mg/dl (1 = true; 0 = false), resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina (1 = yes; 0 = no). As displayed, our proposed model provides 72.05% accuracy, which outstands the result of GMM. We present the results of the model evaluation in Table 2.1. Also, we demonstrate the outcomes of MML approach in Fig. 2.2 that validate our approach for model

selection.

2.4.2 Breast Tissue

According to the World Health Organization (WHO) reports in December 2020 [52], there is a significant change in the global landscape of cancer; breast cancer is now prevalent cancer and is commonly diagnosed with 2.26 million cases. Over the decades, the incidence of diagnosed cases has risen due to daily routines, such as lack of sufficient physical activity and tobacco and alcohol consumption. The strategies presented over these years to reduce this disease's mortality rate and management could not stop most cases because of diagnosis in very late stages. Therefore, early detection is essential in improving treatment outcomes and survival rates. The steps in diagnosis are palpation, mammography, or ultrasound imaging check-up. Analyzing pathological images of breast tissue is required to prevent disease development in case of any doubt about malignancy. The pathologists assess the biopsy tissue regarding microscopic structure. Benign and malignant lesions are differentiated based on dissimilarities in histological characteristics of tissue. Fig. 2.3 illustrates some samples of breast tissues. CAD techniques and integration of machine learning methods in decision-making may reduce false diagnoses and increase efficiency. In this part of our experiment, we tested our method on a publicly available dataset [53] with malignant and benign labels. We applied our model to differentiate the tissues into two clusters, each containing 500 samples. The results in Table 2.2 indicate better performance of our proposed model compared to GMM which provides lower values in four metrics. So, choosing McDBMM has more persuading results. Also, Fig. 2.4 proves that our algorithm was able to find the optimal number of clusters.

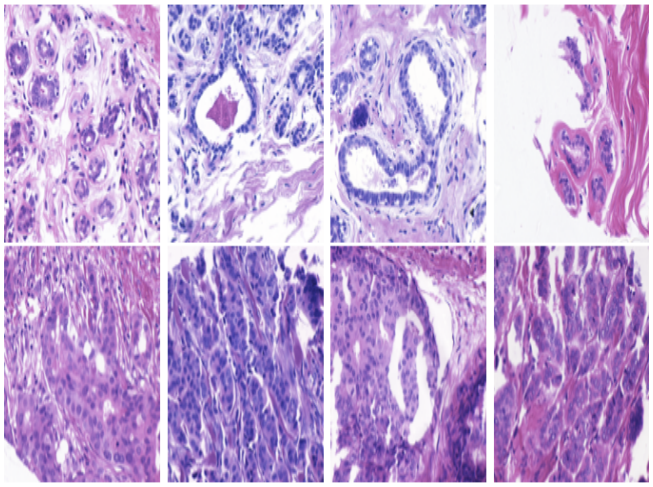

Figure 2.3: Samples of benign and malignant breast tissue. Benign and malignant samples are represented in the first and second rows, respectively.

Table 2.2: Results on Breast Tissue Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
McDBMM	98.33	99.87	96.66	98.30
GMM	73.7	82.11	60.6	69.73

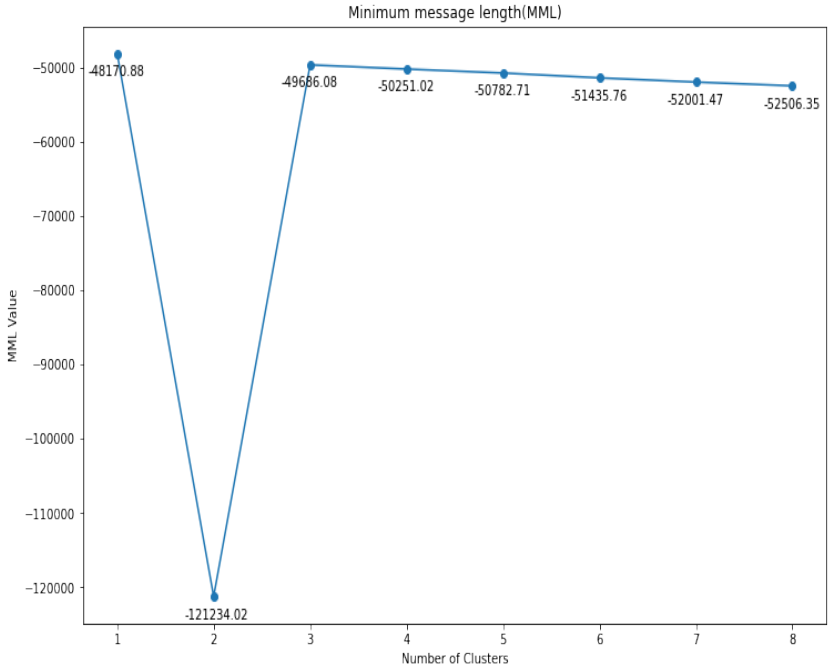

Figure 2.4: Message length plot for the Breast Tissue dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 2.

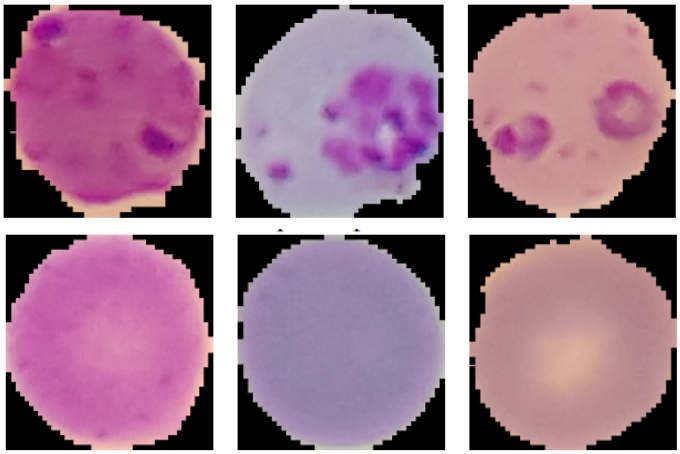

Figure 2.5: Samples of infected cells and uninfected cells. The samples are represented in the first and second rows.

2.4.3 Malaria Detection

Medical areas such as cell biology and screening experiments in diagnosing and predicting illnesses from images obtained by cytological and histological methods are significant applications of cell image clustering. One common challenge in image clustering is that large-scale data prevents manual validations. To face this challenge, image analysis based on machine learning methods is applied to microscopic images to improve the process of cell feature extraction with a more elevated speed. In this section, we studied the performance of our model in malaria detection which is a life-threatening yet preventable and curable disease caused by parasites. Microscopists normally examine blood smears to diagnose and calculate parasitemia by finding the cells infected with malaria. These cells are recognized by the small clot inside the cellular images, whereas uninfected cells are without any clot. Fig. 2.5 represents some image samples of cells. Finding positive cases and verifying all samples in the case of a massive volume of smears is a crucial challenge for humans. In this chapter, we applied our model to a cell image dataset from NIH [1], [54], including 2000 samples. Results in Table 2.3 indicate the robustness of our proposed model compared to GMM and Fig.2.6 indicates the optimal number of clusters.

Table 2.3: Results on Malaria Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
McDBMM	98.30	100	96.60	98.27
GMM	83.3	74.96	100	85.68

[1] `https://ceb.nlm.nih.gov/repositories/malaria-datasets`

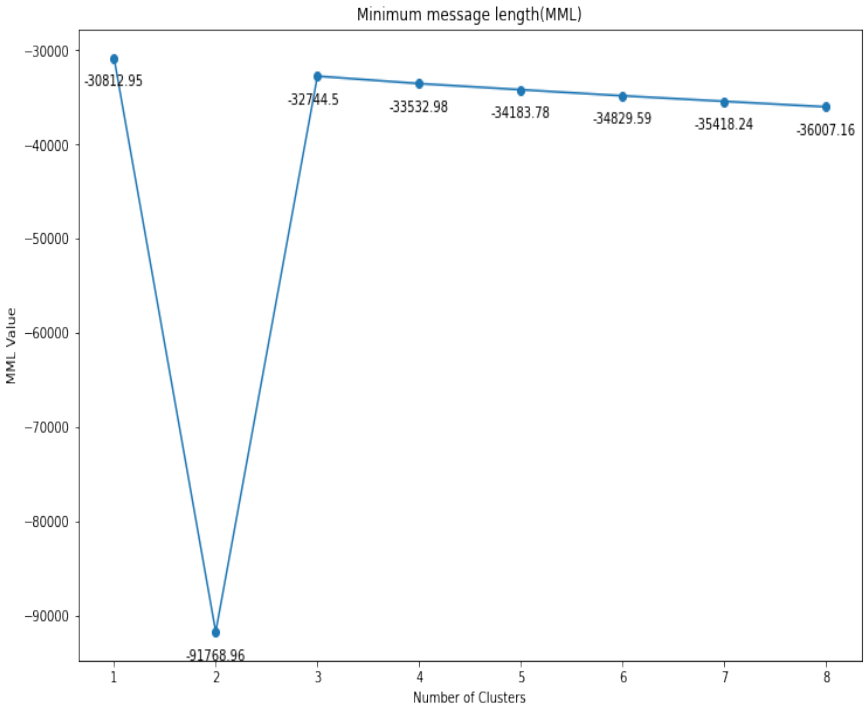

Figure 2.6: Message length plot for the Malaria dataset. The X-axis defines the number of clusters, and the Y-axis defines the value of the message length. According to the plot, The optimal number of clusters is 2.

Chapter 3

A fully Bayesian Inference Approach for Multivariate McDonald's Beta Mixture Model with Feature Selection

As this chapter builds upon the concepts and model specification introduced in Chapter 2, we will not reiterate the definitions and construction of the FMcDBMM. For a comprehensive understanding of the McDonald's Beta distribution and the steps involved in constructing the FMcDBMM, please refer to Chapter 2. In this chapter, we will focus on bayesian framework as the model learning approach with the integration of simultaneous feature selection.

3.1 Feature Saliency

This section introduces the notion of feature saliency, which involves assigning weights to features according to their relevance to the model. While utilizing a large number of features may enhance the model's potential, noise and redundancy can undermine its effectiveness. Thus, by assigning saliency weights, we can improve the model's accuracy, avoid overfitting, and facilitate its interpretation. The equations are initially developed based on a single observation and then extended to apply to all sets of observations for ease of comprehension. Here, we present all observations as a vector $\vec{X}$ where each vector has a D-dimensional representation $\vec{X} = \big(x_1, \ldots, x_D\big)$ and we denote

each dimension as x_d where $d = 1, \ldots, D$. To indicate the relevance of each feature x_d to the j^{th} component of the mixture model, we introduce a set of binary parameters $\mathcal{A} = \{\alpha_{jd}\}$ [55]. If α_{jd} is equal to one, it indicates that the d^{th} feature is relevant to the j^{th} component. If the distribution of the d^{th} feature does not depend on the component labels, it is considered irrelevant. To represent the common density of such irrelevant features, we use $q(. \mid \vec{\Lambda}_d)$. For the purpose of this study, we assume the McDBD as the common density, resulting in the following model:

$$p(\vec{X}_n \mid \vec{w}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) = \sum_{j=1}^{M} w_j \prod_{d=1}^{D} [p((x_{nd} \mid \theta_{jd})]^{\alpha_{jd}} [q(x_{nd} \mid \vec{\Lambda}_d)]^{1-\alpha_{jd}} \tag{3.1}$$

Equation (3.1) defines $\vec{\Lambda}_d$ as the parameter of the common density of the d^{th} feature, where $\vec{\Lambda}_d = (\hat{a_d}, \hat{b_d}, \hat{p_d})$. To describe α_{jd}, we present $\mathcal{P} = \{\rho_{jd}\}$, which is known as the component-based feature saliency. The value of ρ_{jd}, which indicates the degree to which component j^{th} is related to the d^{th} feature, is defined as $p(\alpha_{jd} = 1)$. It is also possible to infer that $p(\alpha_{id} = 0) = 1 - \rho_{jd}$. Thus, we have:

$$p(\alpha_{jd} \mid \rho_{jd}) = \rho_{jd}^{\alpha_{jd}} (1 - \rho_{jd})^{1-\alpha_{jd}} \tag{3.2}$$

3.1.1 Feature Selection in McDonald's Beta Mixture Model

In this section, we will focus on incorporating feature selection into the FMcDBMM using the complete set of observations $\mathcal{X}$. The mixture model can be derived according to (3.1) and (3.2) as follow:

$$p(\vec{X}_n \mid \Delta) = \sum_{j=1}^{M} w_j \prod_{d=1}^{D} (\rho_{jd} p(\ x_{nd} \mid \theta_{jd}) + (1 - \rho_{jd}) q(\ x_{nd} \mid \vec{\Lambda}_d)) \tag{3.3}$$

where

$$\Delta = \{\{w_j\}, \{\theta_{jd}\}, \{\rho_{jd}\}, \{\vec{\Lambda}_d\}\} \tag{3.4}$$

Next, we define a membership vector $\vec{Z}_n = (Z_{n1}, \ldots, Z_{nM})$ of dimension M for each observation $\vec{X}_n$, where $Z_{nj} = 1$ indicates that $\vec{X}_n$ belongs to component j, and $Z_{nj} = 0$ otherwise. Thus, we can consider a set of membership vectors for $\mathcal{X}$ defined by $\mathcal{Z} = (Z_1, \ldots, Z_N)$ and have a complete form of data as $(\mathcal{X}, \mathcal{Z})$ which follows $p(\mathcal{X}, \mathcal{Z} \mid \Delta)$. The density of the complete form of data can be defined as follows:

$$p(\mathcal{X}, \mathcal{Z} \mid \Delta) = \tag{3.5}$$
$$\prod_{n=1}^{N}\prod_{j=1}^{M}[w_j \prod_{d=1}^{D}(\rho_{jd}p(x_{nd} \mid \theta_{jd}) + (1-\rho_{jd})q(x_{nd} \mid \vec{\Lambda}_d))]^{Z_{nj}}$$

By considering the missing multinomial variable $\hat{Z}_{nj}$ for each $\vec{X}_n$, such that $\vec{Z}_n \sim \mathcal{M}(1; \hat{Z}_{n1}, \ldots, \hat{Z}_{nM})$, we have:

$$\hat{Z}_{nj} = \frac{p\left(\vec{X}_n|\vec{\theta_j}\right) w_j}{\sum_{j=1}^{M} p\left(\vec{X}_n|\vec{\theta_j}\right) w_j} \tag{3.6}$$

3.2 Model Learning

3.2.1 Bayesian Learning Framework

Estimating model's parameters poses a challenging topic during the learning step of mixture models. While several deterministic and stochastic techniques exist, we present a Bayesian framework for multivariate McDonald's Beta mixture models due to its capability to integrate prior knowledge and assumptions about the model's parameters and the associated uncertainties into the estimation process. Our proposed framework utilizes the Metropolis-Hastings algorithm and Gibbs sampler technique which will be explained as we proceed.

In Bayesian inference, the initial step involves defining the prior and posterior probability distributions. Thus, we will compute the posterior distribution with the help of Bayes' theorem. Given the complete data

$(\mathcal{X}, \mathcal{Z})$, the joint distribution of $p(\mathcal{X}, \mathcal{Z} \mid \Delta)$ and the prior density function $p(\Delta)$, the posterior is defined as follows:

$$p(\Delta \mid \mathcal{X}, \mathcal{Z}) = \frac{p(\mathcal{X}, \mathcal{Z} \mid \Delta)p(\Delta)}{\int p(\mathcal{X}, \mathcal{Z} \mid \Delta)p(\Delta)} \propto p(\mathcal{X}, \mathcal{Z} \mid \Delta)p(\Delta) \tag{3.7}$$

Considering the fact that mixing weight parameters must satisfy the constraints $0 < w_j < 1$ and $\sum_{j=1}^{M} w_j = 1$, a symmetric Dirichlet distribution with parameters $(\zeta_1, \ldots, \zeta_M)$ is often considered to be the optimal choice for establishing a prior distribution.

$$p(\vec{w}) = \frac{\Gamma(\sum_{j=1}^{M} \zeta_j)}{\prod_{j=1}^{M} \Gamma(\zeta_j)} \prod_{j=1}^{M} w_j^{\zeta_j - 1} \tag{3.8}$$

Furthermore, we have:

$$\begin{aligned} p(\mathcal{Z} \mid \vec{w}) &= \prod_{n=1}^{N} p(Z_n \mid \vec{w}) = \prod_{n=1}^{N} w_1^{Z_{n1}}, \ldots, w_M^{Z_{nM}} \\ &= \prod_{n=1}^{N} \prod_{j=1}^{M} w_j^{Z_{nj}} = \prod_{j=1}^{M} w_j^{n_j} \end{aligned} \tag{3.9}$$

where $n_j = \sum_{n=1}^{N} \mathbb{I}_{Z_{nj}=j}$. Therefore, using (3.8) and (3.9) and based on Bayes theorem, the posterior is defined as:

$$\begin{aligned} p(\vec{w} \mid \mathcal{Z}) &= \frac{\Gamma(\sum_{j=1}^{M} \zeta_j)}{\prod_{j=1}^{M} \Gamma(\zeta_j)} \prod_{j=1}^{M} w_j^{\zeta_j - 1} \prod_{j=1}^{M} w_j^{n_j} \\ &= \frac{\Gamma(\sum_{j=1}^{M} \zeta_j)}{\prod_{j=1}^{M} \Gamma(\zeta_j)} \prod_{j=1}^{M} w_j^{\zeta_j + n_j - 1} \propto \mathcal{D}(\zeta_1 + n_1, \ldots, \zeta_M + n_M) \end{aligned} \tag{3.10}$$

where $\mathcal{D}$ is a Dirichlet distribution with $(\zeta_1 + n_1, \ldots, \zeta_M + n_M)$ as it's parameters.

Afterward, we must establish a prior for ρ_{jd}, representing the feature saliency probabilities. As per its characteristics, our preferred initial prior is the Beta distribution.

$$p(\rho_{jd}) = \frac{\Gamma(t_{jd} + \eta_{jd})}{\Gamma(t_{jd})\Gamma(\eta_{jd})} \rho_{jd}^{t_{jd}-1} (1 - \rho_{jd})^{\eta_{jd}-1} \tag{3.11}$$

Based on (3.11), The posterior is defined as below:

$$\begin{aligned} &p(\mathcal{P} \mid \mathcal{X}, \mathcal{Z}, \mathcal{A}) \\ &= \prod_{j=1}^{M} \prod_{d=1}^{D} \text{Beta}(n_{jd}^{*} + t_{jd}, n_j - n_{jd}^{*} + \eta_{jd}) \end{aligned} \tag{3.12}$$

In (3.12), $n_{jd}^{*} = \sum_{n=1}^{N} z_{nj} \phi_{njd}$ and $\phi_{njd} \in \{0, 1\}$ satisfying following conditions:

$$\phi_{njd} = \begin{cases} 1 & h \geq 1 \\ 0 & \text{else} \end{cases}, \qquad h = \frac{\rho_{jd} p(x_{nd} \mid a_{jd}, b_{jd}, p_{jd})}{(1 - \rho_{jd}) q(x_{nd} \mid \hat{a_d}, \hat{b_d}, \hat{p_d}\})}. \tag{3.13}$$

Furthermore, Given the positivity constraints imposed on all of the parameters, the Gamma distribution would be a proper choice for the prior distribution of each parameter. As a result, the prior distributions can be expressed in the form of the following equations:

$$p\left(\vec{a} \mid u, v\right) = \prod_{j=1}^{M} \prod_{d=1}^{D} \frac{v_{jd}^{u_{jd}}}{\Gamma(u_{jd})} a_{jd}^{u_{jd}-1} e^{-v_{jd} a_{jd}} \tag{3.14}$$

$$p\left(\vec{b} \mid r, s\right) = \prod_{j=1}^{M} \prod_{d=1}^{D} \frac{s_{jd}^{r_{jd}}}{\Gamma(r_{jd})} b_{jd}^{r_{jd}-1} e^{-s_{jd} b_{jd}} \tag{3.15}$$

$$p\left(\vec{p} \mid f, g\right) = \prod_{j=1}^{M} \prod_{d=1}^{D} \frac{g_{jd}^{f_{jd}}}{\Gamma(f_{jd})} p_{jd}^{f_{jd}-1} e^{-g_{jd} p_{jd}} \tag{3.16}$$

$$\begin{aligned} &p(\vec{\theta} \mid \vec{u}, \vec{v}, \vec{r}, \vec{s}, \vec{f}, \vec{g}) = p\left(\vec{a} \mid u, v\right) \\ &p\left(\vec{b} \mid r, s\right) \; p\left(\vec{p} \mid f, g\right) \end{aligned} \tag{3.17}$$

$$p\left(\vec{\hat{a}} \mid \hat{u}, \hat{v}\right) = \prod_{d=1}^{D} \frac{\hat{v}_d^{\hat{u}_d}}{\Gamma(\hat{u}_d)} \hat{a}_d^{\hat{u}_d - 1} e^{-\hat{v}_d \hat{a}_d} \tag{3.18}$$

$$p\left(\vec{\hat{b}} \mid \hat{r}, \hat{s}\right) = \prod_{d=1}^{D} \frac{\hat{s}_d^{\hat{r}_d}}{\Gamma(\hat{r}_d)} \hat{b}_d^{\hat{r}_d - 1} e^{-\hat{s}_d \hat{b}_d} \tag{3.19}$$

$$p\left(\vec{\hat{p}} \mid \hat{f}, \hat{g}\right) = \prod_{d=1}^{D} \frac{\hat{g}_d^{\hat{f}_d}}{\Gamma(\hat{f}_d)} \hat{p}_d^{\hat{f}_d - 1} e^{-\hat{g}_d \hat{p}_d} \tag{3.20}$$

$$\begin{aligned} &p(\vec{\Lambda} \mid \vec{u}, \vec{v}, \vec{r}, \vec{s}, \vec{\hat{f}}, \vec{\hat{g}}) = p\left(\vec{\hat{a}} \mid \hat{u}, \hat{v}\right) \\ &p\left(\vec{\hat{b}} \mid \hat{r}, \hat{s}\right)\ p\left(\vec{\hat{p}} \mid \hat{f}, \hat{g}\right) \end{aligned} \tag{3.21}$$

where all the hyper-parameters $\vec{u} = \{u_{jd}\}, \vec{v} = \{v_{jd}\}, \vec{r} = \{r_{jd}\}, \vec{s} = \{s_{jd}\}, \vec{f} = \{f_{jd}\}, \vec{g} = \{g_{jd}\}$ and $\vec{\hat{u}} = \{\hat{u}_d\}, \vec{\hat{v}} = \{\hat{v}_d\}, \vec{\hat{r}} = \{\hat{r}_d\}, \vec{\hat{s}} = \{\hat{s}_d\}, \vec{\hat{f}} = \{\hat{f}_d\}, \vec{\hat{g}} = \{\hat{g}_d\}$ of the above priors are positive.

With conditioning the likelihood on $\mathcal{A}$ and $\mathcal{Z}$ to facilitate the implementation of further Bayesian inference computation, we will have the following:

$$\begin{aligned} &p(\mathcal{X} \mid \mathcal{Z}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) = \\ &\prod_{n=1}^{N} \prod_{d=1}^{D} [p((x_{nd} \mid \theta_{jd})]^{\alpha_{jd}} [q(x_{nd} \mid \vec{\Lambda}_d)]^{1-\alpha_{jd}} \end{aligned} \tag{3.22}$$

The conditional posterior distributions are computed using the priors and likelihood defined above as follows:

$$
\begin{aligned}
p(\vec{\theta_j} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) &\propto p(\vec{\theta_j}) \prod_{z_{nj}=1} p(X_n \mid \mathcal{Z}, \vec{\theta_j}, \vec{\Lambda}, \mathcal{A}) \\
&\propto p(\vec{\theta_j}) \prod_{z_{nj}=1} \prod_{d=1}^{D} [p((x_{nd} \mid \vec{\theta}_{jd})]^{\alpha_{jd}} [q(x_{nd} \mid \vec{\Lambda}_d)]^{1-\alpha_{jd}}] \\
&\propto \prod_{d=1}^{D} [(\frac{v_{jd}^{u_{jd}}}{\Gamma(u_{jd})} a_{jd}^{u_{jd}-1} e^{-v_{jd} a_{jd}}) \\
&\times (\frac{s_{jd}^{r_{jd}}}{\Gamma(r_{jd})} b_{jd}^{r_{jd}-1} e^{-s_{jd} b_{jd}}) \\
&\times (\frac{g_{jd}^{f_{jd}}}{\Gamma(f_{jd})} p_{jd}^{f_{jd}-1} e^{-g_{jd} p_{jd}}) \\
&\times \prod_{z_{nj}=1} \prod_{d=1}^{D} [\frac{p_{jd} {x_{nd}}^{a_{jd} p_{jd}-1} (1 - {x_{nd}}^{p_{jd}})^{b_{jd}-1}}{B(a_{jd}, b_{jd})}]^{\alpha_{jd}} \\
&\times [\frac{\hat{p}_{jd} {x_{nd}}^{\hat{a}_{jd} \hat{p}_{jd}-1} (1 - {x_{nd}}^{\hat{p}_{jd}})^{\hat{b}_{jd}-1}}{B(\hat{a}_{jd}, \hat{b}_{jd})}]^{1-\alpha_{jd}}]
\end{aligned} \tag{3.23}
$$

$$
p(\vec{\Lambda} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) \propto p(\vec{\Lambda}) \prod_{z_{nj}=1} p(X_n \mid \mathcal{Z}, \vec{\theta_j}, \vec{\Lambda}, \mathcal{A}) \tag{3.24}
$$

The process of computing $p(\vec{\Lambda} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A})$ will be the same as $p(\vec{\theta_j} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A})$ which was explained above only with it's own defined parameters.

Once we have computed all our posteriors, the next step is to estimate the parameters of our model. One efficient method to achieve this is using MCMC techniques, and Gibbs sampling is considered one of the most effective approaches. Using Gibbs sampling, we can approximate the model parameters by sequentially deriving them from their posteriors based on earlier approximated values.

To simulate $\vec{\theta_j}$ from its posterior distribution, we utilize the Metropolis-Hastings approach. This approach requires us to specify a proposal distribution which considering the fact that all the parameters are positive, We have used a random walk M-H with the following proposal distribution:

$$
\tilde{\theta}_{jd} \sim \mathcal{LN}(\log(\theta_{jd}^{(t-1)}), \sigma^2) \tag{3.25}
$$

This is considered the first phase in the M-H algorithm. To elaborate, $\mathcal{LN}(\log(\theta_{jd}^{(t-1)}), \sigma^2)$ is the log-normal distribution with mean and variance of $\log(\theta_{jd}^{(t-1)})$ and σ^2 where $d = 1, \ldots, D$.

Afterward, as the second phase of the M-H algorithm, we will establish an acceptance ratio π to determine whether the new samples generated at iteration t are eligible to be accepted or not. The acceptance ratio is a key component in the M-H algorithm, which controls the exploration of the parameter space and the algorithm's convergence. Mathematically, the acceptance ratio is defined as the ratio of the proposed sample's posterior density to the current sample's posterior density.

$$\begin{aligned}\pi &= \frac{p(\tilde{\theta}_j \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} \mathcal{LN}((\theta_{jd}^{(t-1)}) \mid \log(\tilde{\theta}_{jd}), \sigma^2)}{p(\vec{\theta}_j^{(t-1)} \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} \mathcal{LN}(\tilde{\theta}_{jd} \mid \log(\theta_{jd}^{(t-1)}), \sigma^2)} \\ &= \frac{p(\tilde{\theta}_j \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} \tilde{a}_{jd} \tilde{b}_{jd} \tilde{p}_{jd}}{p(\vec{\theta}_j^{(t-1)} \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} a_{jd}^{(t-1)} b_{jd}^{(t-1)} p_{jd}^{(t-1)}}\end{aligned} \tag{3.26}$$

Algorithm 2 Complete Algorithm

1. Initialization
   (a) Apply K-means and initialize the parameters

   Repeat

2. Gibbs Sampling
   (a) Generate $\vec{Z}_n \sim \mathcal{M}(1; \hat{Z}_{n1}, \ldots, \hat{Z}_{nM})$
   (b) Generate $\vec{w}$, $\mathcal{P}$ from (3.10), (3.12)
3. Metropolis-Hastings
   (a) Generate $\tilde{\theta}_{jd}$ from (3.25)
   (b) Compute π from (3.26)

   until
   Convergence

Table 3.1: Results on Lung Cancer Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
FMcDBMM	87.95	88.1	87.96	88.02
GMM	83.33	79.88	80.43	81.59

3.3 Experimental Results

In this section, we evaluated our proposed model for two real-world applications: Lung cancer analysis and Human activity recognition. To evaluate its effectiveness, we compared the performance of our model with GMM. The model's performance was assessed using four standard metrics: accuracy, precision, recall, and F1-score and the min-max method to normalize our datasets.

$$
\begin{aligned}
Accuracy &= \frac{TruePositives + TrueNegatives}{\text{Total number of observations}} \\
Precision &= \frac{TruePositives}{TruePositives + FalsePositives} \\
Recall &= \frac{TruePositives}{TruePositives + FalseNegatives} \\
F1 - score &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned} \tag{3.27}
$$

3.3.1 Lung Cancer Analysis

we applied our model to a dataset of lung histopathological images [56] as our first experiment. This dataset contains 2500 images classified into three varieties of lung cancer: benign, adenocarcinoma, and squamous cell carcinoma. Fig. 4.1 represents some image samples of cells.

Moreover, The results in Table 4.1 indicate better performance of our proposed model with an accuracy of 87.95% compared to GMM with an accuracy of 83.33%. Also, we have illustrated the feature saliencies of eight random features in Fig. 3.2, where according to the chart the eighth feature has the most relevancy among the other seven features across all components.

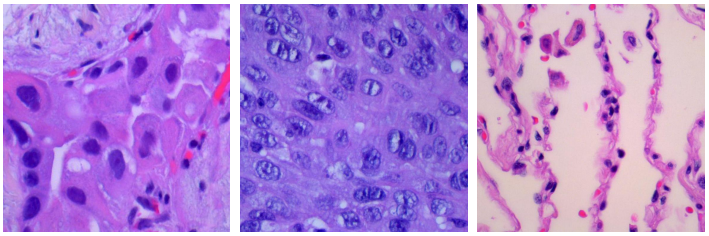

Figure 3.1: Lung Cancer cell image samples

Table 3.2: Results on Human Activity Recognition Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
FMcDBMM	89.27	89.18	89.34	89.25
GMM	86.71	86.39	86.41	86.39

3.3.2 Human Activity Recognition (HAR)

In our second experiment, we use a publicly available dataset with 2220 samples [57] consisting of four activities - laying, sitting, standing, and walking. The data was collected from the activities of 30 participants aged between 19 and 48 while they wore a waist-mounted Samsung Galaxy S II smartphone. Two sensors, an accelerometer, and a gyroscope were embedded in the smartphone to record the data. Note that, the complexity and instability of sensor-based data make it challenging to analyze human activity.

Additionally, attributes including triaxial acceleration derived from the accelerometer (overall acceleration), the approximated body acceleration, and triaxial angular velocity obtained from the gyroscope were utilized. The results presented in Table 4.3 show that our proposed model achieved a promising performance compared to GMM. Correspondingly, Fig. 3.3 illustrates the feature saliency among components for eight sample features.

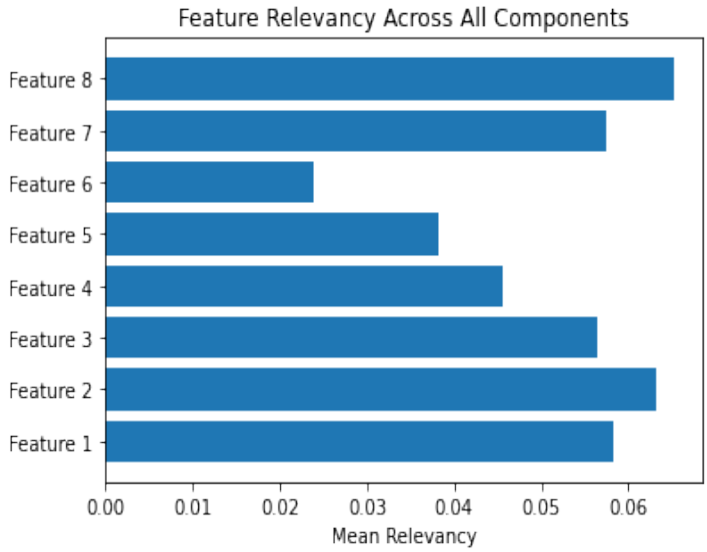

Figure 3.2: Feature relevancy of eight random features in Lung cancer dataset across all components

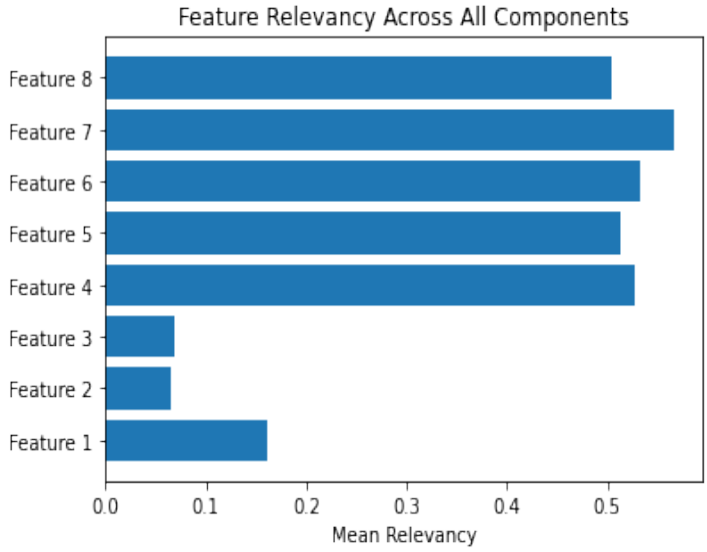

Figure 3.3: Feature relevancy of eight random features in Human activity recognition dataset across all components

Chapter 4

Bayesian Inference in Infinite Multivariate McDonald's Beta Mixture model

As this chapter is an extension of the concepts and methods introduced in Chapter 3, we will not provide a detailed account of model specifications and feature selection process here. For a thorough understanding of model specification as well as the feature selection employed, we kindly direct the reader to Chapter 3.

In the present chapter, our primary focus will be on expanding the finite McDonald's Beta Mixture Model into an infinite mixture model, discussing its potential implications and applications within the medical data analysis and human activity recognition domain.

4.1 Model Learning

4.1.1 Bayesian Learning Framework

As discussed in Chapter 3where we delved into the details of the Bayesian framework for model learning, Given the complete data $(\mathcal{X}, \mathcal{Z})$, the joint distribution of $p(\mathcal{X}, \mathcal{Z} \mid \Delta)$ and the prior density function $p(\Delta)$, we will define the posterior distribution for $\mathcal{P}$ as below:

$$p(\mathcal{P} \mid \mathcal{X}, \mathcal{Z}, \mathcal{A}) = \prod_{j=1}^{M} \prod_{d=1}^{D} \text{Beta}(n^*_{jd} + t_{jd}, n_j - n^*_{jd} + \eta_{jd}) \tag{4.1}$$

where $n^*_{jd} = \sum_{n=1}^{N} z_{nj}\phi_{njd}$ and $\phi_{njd} \in \{0, 1\}$ meets the following conditions:

$$\phi_{njd} = \{ \begin{array}{ll} 1 & h \geq 1 \\ 0 & \text{else} \end{array}, \quad h = \frac{\rho_{jd} p(x_{nd} \mid a_{jd}, b_{jd}, p_{jd})}{(1-\rho_{jd}) q(x_{nd} \mid \hat{a_d}, \hat{b_d}, \hat{p_d}\})}. \tag{4.2}$$

As mentioned before in chapter 3, we calculate the model parameter posteriors as follows:

$$\begin{aligned} p(\vec{\theta_j} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) \propto \prod_{d=1}^{D} [(\frac{v_{jd}^{u_{jd}}}{\Gamma(u_{jd})} a_{jd}^{u_{jd}-1} e^{-v_{jd}a_{jd}}) \times (\frac{s_{jd}^{r_{jd}}}{\Gamma(r_{jd})} b_{jd}^{r_{jd}-1} e^{-s_{jd}b_{jd}}) \\ \times (\frac{g_{jd}^{f_{jd}}}{\Gamma(f_{jd})} p_{jd}^{f_{jd}-1} e^{-g_{jd}p_{jd}}) \times \prod_{z_{n,j}=1} \prod_{d=1}^{D} [\frac{p_{jd} {x_{nd}}^{a_{jd}p_{jd}-1} (1-{x_{nd}}^{p_{jd}})^{b_{jd}-1}}{B(a_{jd}, b_{jd})}]^{\alpha_{jd}} \\ \times [\frac{\hat{p}_{jd} {x_{nd}}^{\hat{a}_{jd}\hat{p}_{jd}-1} (1-{x_{nd}}^{\hat{p}_{jd}})^{\hat{b}_{jd}-1}}{B(\hat{a}_{jd}, \hat{b}_{jd})}]^{1-\alpha_{jd}}] \end{aligned} \tag{4.3}$$

$$p(\vec{\Lambda} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A}) \propto p(\vec{\Lambda}) \prod_{z_{nj}=1} p(X_n \mid \mathcal{Z}, \vec{\theta_j}, \vec{\Lambda}, \mathcal{A}) \tag{4.4}$$

$p(\vec{\Lambda} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A})$ will be computed in the same way as $p(\vec{\theta_j} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}, \vec{\Lambda}, \mathcal{A})$.

4.1.2 Extension to Infinite Mixture Model

Determining the number of components M to accurately represent data is essential but difficult. As the need to set M beforehand is a major drawback, researchers have suggested nonparametric Bayesian techniques, which can automatically figure out the number of clusters and expand them indefinitely based on a specific choice of prior for mixing weights [58]. Unlike finite mixture models, where each vector $\vec{X_n}$ is derived from one of M undefined McDBD, we present that a Dirichlet process of McDBD to model our data. In the following, we will demonstrate the fundamentals of the Dirichlet process mixture model and its ability to create or eliminate components. consider a symmetric Dirichlet with a concentration parameter $\frac{\tau}{M}$ as the prior for mixing weights:

$$p(\vec{w} \mid \tau) = \frac{\Gamma(\tau)}{\prod_{j=1}^{M} \Gamma(\frac{\tau}{M})} \prod_{j=1}^{M} w_j^{\frac{\tau}{M}-1} \tag{4.5}$$

where $\vec{w}$ is the vector of mixing weights defined by $(w_1, \ldots, w_M) : \sum_{j=1}^{M-1} w_j < 1$. In addition, for Z_n as the latent variable to show which cluster does X_n belongs to, such that $w_j = p(Z_n = j), j = 1, \ldots, M$ we have the following:

$$p(\mathcal{Z} \mid \vec{w}) = \prod_{j=1}^{M} {w_j}^{n_j} \tag{4.6}$$

where $n_j = \sum_{n=1}^{N} \mathbb{I}_{Z_{nj}=1}$. Since the Dirichlet distribution is a conjugate prior for the multinomial distribution, we can calculate the prior distribution for $\mathcal{Z}$ by integrating out the mixing proportions vector $\vec{w}$ as follows:

$$p(\mathcal{Z} \mid \tau) = \int_{\vec{w}} p(Z \mid \vec{w}) p(\vec{w} \mid \tau) d\vec{w} = \frac{\Gamma(\tau)}{\Gamma(N+\tau)} \prod_{j=1}^{M} \frac{\Gamma(\frac{\tau}{M} + n_j)}{\Gamma(\frac{\tau}{M})} \tag{4.7}$$

Therefore using (4.5) to (4.7), we obtain:

$$p(\vec{w} \mid \mathcal{Z}, \tau) = \frac{\Gamma(\tau + N)}{\prod_{j=1}^{M} \Gamma(\frac{\tau}{M} + n_j)} \prod_{j=1}^{M} w_j^{n_j + \frac{\tau}{M} - 1} \propto \mathcal{D}(n_1 + \frac{\tau}{M}, \ldots, n_M + \frac{\tau}{M}) \tag{4.8}$$

where $\mathcal{D}$ is a Dirichlet distribution with parameters $(n_1 + \frac{\tau}{M}, \ldots, n_M + \frac{\tau}{M})$. According to [59], the conditional prior for a single indicator is defined as below:

$$p(Z_{nj} = 1 \mid \tau, \mathcal{Z}_{-n}) = \frac{n_{-nj} + \frac{\tau}{M}}{N - 1 + \tau} \tag{4.9}$$

where $\mathcal{Z}_{-n}$ is $\mathcal{Z}$ excluding Z_n and n_{-nj} is the number of observations excluding $\vec{X}_n$ which belongs to cluster j. To tackle the model's complexity challenges, we will suppose $M \to \infty$ and by applying that on (4.9), we have the following [59]:

$$p(Z_{nj} = 1 \mid n; \mathcal{Z}_{-n}) = \begin{cases} \frac{n_{-nj}}{N-1+\tau} & \text{if } n_{-nj} > 0 \quad (j \in \mathcal{R}) \\ \frac{\tau}{N-1+\tau} & \text{if } n_{-nj} = 0 \quad (j \in \mathcal{U}) \end{cases} \tag{4.10}$$

Note that $\mathcal{R}$ and $\mathcal{U}$ indicate the sets of represented and unrepresented components. It is worth noting that the conditional prior distribution for the members of R is dependent on the number of observations assigned to the component, whereas, for the members of U, it only depends on the parameters

τ and N [60]. Therefore having (4.10) as the priors, we present the conditional posteriors [61]:

$$p(Z_{nj} = 1 \mid \vec{\theta_j}, \vec{\Lambda}, \alpha_j, \tau; \mathcal{Z}_{-n}) = \begin{cases} \frac{n_{-nj}}{N-1+\tau} p(\vec{X_n} \mid \vec{\theta_j}, \vec{\Lambda}, \alpha_j) & \text{if } j \in \mathcal{R} \\ \int \frac{\tau p(\vec{X_n} \mid \vec{\theta_j}, \vec{\Lambda}, \alpha_j) p(\vec{\theta_j}, \vec{\Lambda})}{N-1+\tau} d\vec{a_j} d\vec{b_j} d\vec{p_j} d\vec{\Lambda} & \text{if } j \in \mathcal{U} \end{cases} \tag{4.11}$$

Equation (4.11) represents a DP mixture model [62]. Each observation is assigned to a cluster based on a set of mixing proportions, with the number of clusters determined automatically from the data. One of the key advantages of this model is its ability to adapt to the data and generate new clusters as needed, avoiding overfitting and allowing for a flexible representation of the data. If an observation is assigned to an unrepresented cluster, a new cluster is generated to accommodate it, while a represented cluster may become unrepresented if all its observations are assigned to other clusters during the sampling process.

4.2 Algorithm Overview

To estimate the parameters of our mixture model as our final step, we have used the M-H algorithm and Gibbs sampler technique [63], which are widely used in Bayesian inference to sample from complex posterior distributions and avoid direct sampling. For the choice of the proposal distribution, we have used a random walk M-H with the following proposal distribution:$\tilde{\theta}_{jd} \sim \mathcal{LN}(\log(\theta_{jd}^{(t-1)}), \sigma^2)$, where $\mathcal{LN}(\log(\theta_{jd}^{(t-1)}), \sigma^2)$ is the log-normal distribution with mean and variance of $\log(\theta_{jd}^{(t-1)})$ and σ^2 where $d = 1, \ldots, D$.

To summarize the algorithm, first, we initialize the algorithm by assigning all observations to the same cluster. We then generate the vector $\vec{Z_n}$ and update the number of represented clusters based on the generated vector. Since the integral in equation (4.11) is not analytically tractable, we employ the technique proposed in [64] to approximate it and enable sampling from the vector $\vec{Z_n}$.

Algorithm 3 Nonparametric Bayesian learning of IMcDBMM

1. Initialization
2. **Repeat**
3. Generate $\vec{Z_n}$ from (4.11) and then update n_j.
4. Update the number of represented components.
5. Update the mixing weights for the represented components by $w_R = \frac{n_j}{N+\tau}$.
6. Update the mixing weights for the unrepresented components by $w_U = \frac{\tau}{N+\tau}$.
7. Generate the model parameters from (4.3) using M-H algorithm.
8. **until** Convergence

4.3 Experimental Results

We evaluate our proposed model on two real-world applications: lung cancer analysis and HAR. We compared its performance with the widely-used GMM and FMcDBMM using accuracy, precision, recall, and F1-score metrics to assess its effectiveness.

4.3.1 Lung Cancer Analysis

We conducted our first experiment on a dataset of lung cancer images [56], which contains 2500 images classified into three categories: benign, adenocarcinoma, and squamous cell carcinoma. Examples of cell image samples are shown in fig.4.1. We compared the performance of our proposed model with GMM using the mentioned metrics. The results in Table 1 indicate that our model achieved better accuracy of 90.84% than GMM and FMcDBMM. In addition, we analyzed the feature saliencies of eight random features, as shown in Table 4.2. The analysis indicated that the eighth feature was the most relevant among the other seven features across all components.

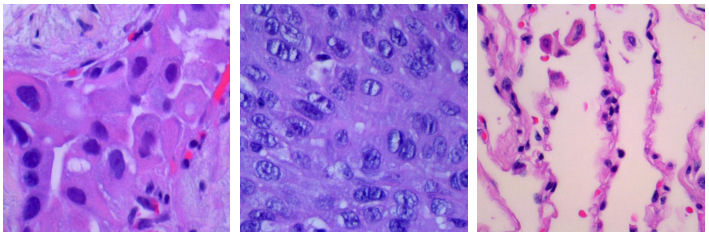

Figure 4.1: Sample of each type of lung cancer images

Table 4.1: Results on Lung Cancer Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
IMcDBMM	90.84	91.02	90.88	90.94
FMcDBMM	87.95	88.1	87.96	88.02
GMM	83.33	79.88	80.43	81.59

Table 4.2: Feature Relevancy Across All Components for Lung Cancer dataset

Feature	F1	F2	F3	F4	F5	F6	F7	F8
Relevancy	0.1092	0.1142	0.1074	0.0965	0.0891	0.0747	0.1083	0.1163

4.3.2 Human Activity Recognition (HAR)

As our second experiment, we used a publicly available dataset of 2220 samples [57] collected from 30 participants aged between 19 and 48 performing four activities (laying, sitting, standing, and walking) while wearing a waist-mounted Samsung Galaxy S II smartphone with two embedded sensors. We employed features such as triaxial acceleration, the estimated body acceleration, and triaxial angular velocity to analyze the data. Our proposed model achieved promising performance compared to GMM and FMcDBMM, as shown in Table 4.3. Additionally, we analyzed the feature saliency among components for eight sample features, as illustrated in Table 4.4.

Table 4.4: Feature Relevancy Across All Components for HAR dataset

Feature	F1	F2	F3	F4	F5	F6	F7	F8
Relevancy	0.5275	0.4327	0.4350	0.8949	0.8794	0.9001	0.9340	0.8709

Table 4.3: Results on HAR Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
IMcDBMM	93.14	91.32	91.27	91.29
FMcDBMM	89.27	89.18	89.34	89.25
GMM	86.71	86.39	86.41	86.39

Chapter 5

Conclusion

In this thesis, we developed various unsupervised methods and applied them to medical problems with a primary aim of providing potent alternatives to commonly utilized models such as the Gaussian Mixture. Our models were grounded on a novel distribution, the multivariate McDonald's Beta distribution. This choice was driven by the reality that the assumption of Gaussianity, which many approaches rely on, often does not hold for many datasets across different fields of science and various real-world applications. The multivariate McDonald's Beta distribution offers a more flexible alternative, capable of modeling symmetric, asymmetric, and skewed data.

First we proposed a novel finite mixture model based on McDonald's Beta distribution, showcasing its flexibility and suitability for real-world applications. Employing maximum likelihood via the expectation maximization algorithm, we managed to estimate the parameters of our model, with the minimum message length (MML) criterion assisting in determining the optimal number of clusters. The superior performance of our method, when tested on multiple medical datasets and compared with the Gaussian Mixture Model (GMM), indicated the robustness and viability of our model as an alternative to traditional methods.

We then focused on a Bayesian learning framework, incorporating Markov Chain Monte Carlo techniques and simultaneous feature selection. This model was rigorously tested on real-world applications like lung cancer image analysis and human activity recognition. The experimental results showed our model's superior performance over GMM, laying the groundwork for future expansion to an infinite mixture model.

Lastly, we expanded our finite model to an infinite one, by using a nonparametric Bayesian framework. This extension allowed us to determine the best number of components, adding to the model's flexibility. Our proposed framework outperformed both the GMM and the FMcDBMM in real-world applications, highlighting its effectiveness.

In conclusion, The results from our real-world experiments demonstrate the robustness and flexibility of our proposed models.

Future works could be devoted to integrate the proposed generative models within discriminative ones or to integrate them within deep learning techniques.

Appendix

Determinant of the Fisher information

According to the method presented by [46], we suppose that data samples $X_j = (\vec{X}_s, \ldots, \vec{X}_{s+n_j-1})$ are assigned to the jth cluster, where $s \leq N$ and n_j is the total number of data samples assigned to cluster j. The negative of the second derivative of complete log-likelihood defines $F(\vec{\theta_j})$.

$$
\begin{aligned}
-\log p(\mathcal{X} \mid \Theta) &= -\log\Big(\prod_{n=s}^{s+n_j-1} p(\vec{X} \mid \vec{\theta}_M)\Big) \\
&= -\Big(\sum_{n=s}^{s+n_j-1} \log p(\vec{X} \mid \vec{\theta}_M)\Big)
\end{aligned} \tag{1}
$$

We calculate second and mixed derivative with respect to the parameters a_{jd}, b_{jd}, p_{jd}:

$$
-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd}^2} = -n_j(\psi'(a_{jd} + b_{jd}) - \psi'(a_{jd})) \tag{2}
$$

$$
-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} a_{jd_t}} = 0, d_s \neq d_t \tag{3}
$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd} b_{jd}} = -n_j(\psi'(a_{jd} + b_{jd})) \tag{4}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} b_{jd_t}} = 0, d_s \neq d_t \tag{5}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd} p_{jd}} = -(\sum_{n=1}^{N} \log x_{nd}) \tag{6}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} p_{jd_t}} = 0, d_s \neq d_t \tag{7}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd} a_{jd}} = -n_j(\psi'(a_{jd} + b_{jd})) \tag{8}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} a_{jd_t}} = 0, d_s \neq d_t \tag{9}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd}^2} = -n_j(\psi'(a_{jd} + b_{jd}) - \psi'(b_{jd})) \tag{10}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} b_{jd_t}} = 0, d_s \neq d_t \tag{11}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd} p_{jd}} = -\sum_{n=1}^{N} \frac{\log(x_{nd}) x_{nd}^{p_{jd}}}{x_{nd}^{p_{jd}} - 1} \tag{12}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} p_{jd_t}} = 0, d_s \neq d_t \tag{13}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd} a_{jd}} = -(\sum_{n=1}^{N} \log x_{nd}) \tag{14}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd_s} a_{jd_t}} = 0, d_s \neq d_t \tag{15}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd} b_{jd}} = -\sum_{n=1}^{N} \frac{\log(x_{nd}) x_{nd}^{p_{jd}}}{x_{nd}^{p_{jd}} - 1} \tag{16}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd_s} b_{jd_t}} = 0, d_s \neq d_t \tag{17}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd}^2} = -\sum_{n=1}^{N} \frac{(1 - b_{jd}) x_{nd}^{p_{jd}} \{\log(x_{nd})\}^2}{(1 - x_{nd}^{p_{jd}})^2} - \frac{1}{p_{jd}^2} \tag{18}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial p_{jd_s} p_{jd_t}} = 0, d_s \neq d_t \tag{19}$$

The $F(\vec{\theta_j})$ is a $3D$ by $3D$ matrix as shown below:

$$F_j = \begin{bmatrix} F_{(a_{jd}, a_{jd})} & F_{(a_{jd}, b_{jd})} & F_{(a_{jd}, p_{jd})} \\ F_{(b_{jd}, a_{jd})} & F_{(b_{jd}, b_{jd})} & F_{(b_{jd}, p_{jd})} \\ F_{(p_{jd}, a_{jd})} & F_{(p_{jd}, b_{jd})} & F_{(p_{jd}, p_{jd})} \end{bmatrix} \tag{20}$$

Finally, we calculated the determinant of this block matrix, using the method presented in [65].

List of References

[1] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[2] N. Bouguila. Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Transactions on Knowledge and Data Engineering*, 24(12):2184–2202, 2012.

[3] W. Chen and G. Feng. Spectral clustering with discriminant cuts. *Knowledge-Based Systems*, 28:27–37, 2012.

[4] N. Bouguila. On multivariate binary data clustering and feature weighting. *Comput. Stat. Data Anal.*, 54(1):120–134, 2010.

[5] G. J. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[6] W. Fan, N. Bouguila, and D. Ziou. Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1670–1685, 2013.

[7] B. S. Oboh and N. 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, 2017.

[8] N. Bouguila and W. Fan. *Mixture models and applications*. Springer, 2020.

[9] N. Bouguila, Kh. Almakadmeh, and S. Boutemedjet. A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. *Expert Syst. Appl.*, 39(7):6641–6656, 2012.

[10] I. Channoufi, S. Bourouis, N. Bouguila, and K. 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.

[11] C. Liu, H. Li, K. Fu, F. Zhang, M. Datcu, and W.J. Emery. Bayesian estimation of generalized gamma mixture model based on variational em algorithm. *Pattern Recognition*, 87:269–284, 2019.

[12] N. Manouchehri and N. Bouguila. Stochastic expectation propagation learning of infinite multivariate beta mixture models for human tissue analysis. In *IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2021.

[13] M. Amirkhani, N. Manouchehri, and N. Bouguila. Fully bayesian learning of multivariate beta mixture models. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 120–127. IEEE, 2020.

[14] N. Manouchehri, N. Bouguila, and W. Fan. Nonparametric variational learning of multivariate beta mixture models in medical applications. *International Journal of Imaging Systems and Technology*, 31(1):128–140, 2021.

[15] S. Boutemedjet, D. Ziou, and N. Bouguila. Model-based subspace clustering of non-gaussian data. *Neurocomputing*, 73(10-12):1730–1739, 2010.

[16] W. Fan and N. Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1323–1329, 2013.

[17] W. Fan and N. Bouguila. Variational learning of finite beta-liouville mixture models using component splitting. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2013.

[18] C. Hu, W. Fan, J-X. Du, and N. Bouguila. A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, 333:110–123, 2019.

[19] N. Bouguila. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognit. Lett.*, 33(2):103–110, 2012.

[20] W. Fan and N. Bouguila. Variational learning for dirichlet process mixtures of dirichlet distributions and applications. *Multimedia Tools and Applications*, pages 1–18, 2012. In press.

[21] N. Bouguila and D. Ziou. Dirichlet-based probability model applied to human skin detection [image skin detection]. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–521. IEEE, 2004.

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

[23] W. Fan and N. Bouguila. Online variational learning of generalized dirichlet mixture models with feature selection. *Neurocomputing*, 2013. In press.

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

[25] W. Fan, H. Sallay, and N. Bouguila. Online learning of hierarchical pitman–yor process mixture of generalized dirichlet distributions with feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2048–2061, 2017.

[26] G. Aryal and S. Nadarajah. Information matrix for beta distributions. *Serdica Mathematical Journal*, 30(4):513–526, 2004.

[27] Z. Lu and H. H. S. Ip. Generalized competitive learning of gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 39(4):901–909, 2009.

[28] T. Elguebaly and N. Bouguila. Bayesian learning of generalized gaussian mixture models on biomedical images. In Friedhelm Schwenker and Neamat El Gayar, editors, *Artificial Neural Networks in Pattern Recognition, 4th IAPR TC3 Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings*, volume 5998 of *Lecture Notes in Computer Science*, pages 207–218. Springer, 2010.

[29] W. M. Bolstad and J. M. Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

[30] T. Elguebaly and N. Bouguila. Bayesian learning of finite generalized gaussian mixture models on images. *Signal Process.*, 91(4):801–820, 2011.

[31] N. Bouguila and T. Elguebaly. A fully bayesian model based on reversible jump MCMC and finite beta mixtures for clustering. *Expert Syst. Appl.*, 39(5):5946–5959, 2012.

[32] N. Bouguila and D. Ziou. A dirichlet process mixture of dirichlet distributions for classification and prediction. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pages 297–302, 2008.

[33] N. Bouguila and D. Ziou. A countably infinite mixture model for clustering and feature selection. *Knowl. Inf. Syst.*, 33(2):351–370, 2012.

[34] W. Fan, N. Bouguila, J. Du, and X. Liu. Axially symmetric data clustering through dirichlet process mixture models of watson distributions. *IEEE Transactions on Neural Networks and Learning Systems*, 30(6):1683–1694, 2019.

[35] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1154–1166, 2004.

[36] S. Boutemedjet, D. Ziou, and N. Bouguila. Unsupervised feature selection for accurate recommendation of high-dimensional image data. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 177–184. Curran Associates, Inc., 2007.

[37] N. Elguebaly, T. Bouguila. Simultaneous bayesian clustering and feature selection using rjmcmc-based learning of finite generalized dirichlet mixture models. *Signal Process.*, 93(6):1531–1546, 2013.

[38] H. Lian. Sparse bayesian hierarchical modeling of high-dimensional clustering problems. *Journal of Multivariate Analysis*, 101(7):1728–1737, 2010.

[39] K. Ketabchi, N. Manouchehri, and N. Bouguila. Fully bayesian libby-novick beta mixture model with feature selection. In *2022 IEEE International Conference on Industrial Technology (ICIT)*, pages 1–6. IEEE, 2022.

[40] N. Bouguila and D. Ziou. Online clustering via finite mixtures of dirichlet and minimum message length. *Eng. Appl. Artif. Intell.*, 19(4):371–379, 2006.

[41] N. Bouguila and D. Ziou. Mml-based approach for finite dirichlet mixture estimation and selection. In Petra Perner and Atsushi Imiya, editors, *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, volume 3587 of *Lecture Notes in Computer Science*, pages 42–51. Springer, 2005.

[42] D. Forouzanfar, N. Manouchehri, and N. Bouguila. Finite multivariate mcdonald's beta mixture model learning approach in medical applications. In *Proc. SAC 2023*, 2023.

[43] D. Forouzanfar, N. Manouchehri, and N. Bouguila. A fully bayesian inference approach for multivariate mcdonald's beta mixture model with feature selection. In *Proceedings of the 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2023.

[44] D. Forouzanfar, N. Manouchehri, and N. Bouguila. Bayesian inference in infinite multivariate mcdonald's beta mixture model. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, 2023.

[45] C. S. Wallace and D. L. Dowe. Mml clustering of multi-state, poisson, von mises circular and gaussian distributions. *Statistics and Computing*, 10(1):73–83, 2000.

[46] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

[47] R. Baxter and J. Oliver. Finding overlapping components with mml. *Proc. Third ASTED Conf. Artificial Intelligence and Applications*, 10(1):5–16, 2000.

[48] N. Bouguila and D. Ziou. 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.

[49] W. Jefferys and J. Berger. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

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

[51] A. A. Awan. Heart disease patients dataset, 2020. Dataset Link.

[52] World Health Organization. Breast cancer now most common form of cancer: WHO taking action, 2021. Dataset Link.

[53] M. Paul. Breast histopathology images dataset, 2017. Dataset Link.

[54] Arunava. Malaria cell images dataset, 2018. Dataset Link.

[55] X. Hong, H. Li, P. Miller, J. Zhou, L. Li, D. Crookes, Y. Lu, X. Li, and H. Zhou. Component-based feature saliency for clustering. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):882–896, 2019.

[56] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000), 2019. Dataset Link.

[57] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones, 2013. Dataset Link.

[58] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1), mar 2006.

[59] M. N. Radford. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, jun 2000.

[60] G. Xu and W. Zhou. Bayesian inference for mixture models with an unknown number of components using reversible jump mcmc. *Computational Statistics & Data Analysis*, 132:1–15, 2019.

[61] N. Bouguila and D. Ziou. A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, 21(1):107–122, jan 2010.

[62] N. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian Nonparametrics*. Cambridge University Press, apr 2010.

[63] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[64] C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems*, pages 554–560, 2000.

[65] P. Powell. Calculating determinants of block matrices. *arXiv preprint arXiv:1112.4379*, 2011.