Generative learning models and applications in healthcare

Narges Manouchehri

A Doctoral Thesis
in the Department of
Concordia Institute for Information Systems
Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
"Information and Systems Engineering" at
Concordia University
Montréal, Québec, Canada

June 2022

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared:

By: Narges Manouchehri
Entitled: Generative learning models and applications in healthcare

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

Doctor of Philosophy of **"Information and Systems Engineering"**

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. Kudret Demirli	———————	Chair
Dr. Hussein Yahia	———————	External Examiner
Dr. Farjad Shadmehri	———————	External Examiner
Dr. Chun Wang	———————	Internal Examiner
Dr. Fereshteh Mafakheri	———————	Internal Examiner
Dr. Nizar Bouguila	———————	Supervisor

Approved by ———————————————————

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

June 6, 2022 ———————————————————

Dr. Mourad Debbabi
Dean of Faculty of Engineering and Computer Science

Abstract

Generative learning models and applications in healthcare

Narges Manouchehri, Ph.D.
Concordia University, 2022

Analysis of medical data and making precise decisions by machine learning is emerging as a hot topic in healthcare. The ultimate goal of using these techniques is to transform data into actionable knowledge for supporting clinicians and improving patients' quality of life. To assist health professionals in making precise decisions, clustering is among the most applied methods. This approach aims to stratify patients into meaningful groups based on their similarities in medical data spaces such as images, signals, and medical records. Among all clustering approaches, mixture models have been widely used by researchers in different fields due to their substantial flexibility to explain the data. Traditionally, Gaussian mixture models have been applied in real-world applications but data are not Gaussian in many fields.

In this thesis, we proposed novel clustering approaches based on finite and infinite mixture models. Our mixture models are developed based on a new distribution called multivariate Beta distribution which demonstrated lots of flexibility to fit data of different shapes.

We paired our models with capable learning methods such as variational inference and expectation propagation. These learning methods determine the correct number of mixture components and estimate model's parameters which are two known challenges while fitting mixture models.

Moreover, we modeled sequential data and developed a novel version of the hidden Markov model, namely multivariate Beta-based hidden Markov model, and extended it to a nonparametric model to increase its flexibility.

All developed models are evaluated on real medical applications including medical images and signals. The outcomes demonstrated that our proposed models outperform similar alternatives.

Acknowledgments

I would like to express my deepest gratitude to my wonderful supervisor, Professor Nizar Bouguila who opened the door of the professional research and academic world to me, trusted me, and provided me with the chance to continue my graduate studies under his supervision. It was a great honor to work under his guidance and learn from him. I am also so grateful because of his valuable support of my scholarship applications, nominating me as a student ambassador and letting me have an internship at Ericsson company.

I am also thankful to the members of my Ph.D. committee for their insightful comments and advice.

I would like to thank following organizations for their financial support:
* Fonds de recherche du Québec-Nature et technologies (FQRNT), Ph.D. scholarship for 4 years (2020-24).
* Natural Sciences and Engineering Research Council of Canada, Ph.D. scholarship for 3 years (2021-24).
* Natural Sciences and Engineering Research Council of Canada, Postdoctoral fellowship scholarship for my future research at Karolinska Institute in Sweden for two years (2022-24).
* Concordia University for Carolyn Renaud Teaching Assistantship award in 2020, Professor Hugh McQueen Award of Excellence in 2021, conference and exposition awards in 2020 and 2021.
* Ericsson Global Artificial Intelligence Accelerator (GAIA) and Mitacs (accelerate fellowship program) in 2021.

My sincere appreciation to the following Professors, team leaders and advisors who supported me in my Ph.D. journey:
* Dr. Narsis Aftab Kiani at Karolinska Institute: She dedicated lots of her valuable time to brainstorming and preparing proposals for FRQNT and NSERC Ph.D. scholarships as well as post-doctoral NSERC fellowship scholarship. Receiving such prestigious scholarships from recognized national and provincial organizations became possible thanks to her wonderful and professional guidance. It is a great honor to continue my research as a post-doctoral research fellow under her supervision at Karolinska Institute, which is among the top 10 medical universities in the world (ranked 6^{th} in 2021).

* Dr. Michael Verwey at Concordia University, School of Graduate Studies: During the last couple of years, he supported me by his great advice and instructions in scholarship applications and professional life.
* Dr. Elham Hedayati at Karolinska Institute: She supported my post-doctoral NSERC application.

I would like to express my heartiest gratitude to Professor Ayda Basyouni. She is one of the greatest Professors I have ever had in my life.
It is noteworthy to mention my gratitude to Ms. Silvie Pasquarelli and Mireille Wahba who supported me with their kind help and advice.

I would also like to thank my colleagues at Shared Services Canada. Special thanks to President Sony Perron as Deputy Minister Champion of Concordia University, Mr. Matt Davies as Chief Technology Officer and Ms. Shannon Archibald as Deputy Chief Technology Officer for their valuable advice. I owe profound sincere thanks to Ms. Denise Gomes as Director General of Strategic Engagement and Service Delivery directorate. My sincere gratitude and appreciation to cloud services team leader Mr. Frederic Chakra because of his valuable supports and considerations.

I take this opportunity and express my profound respect and appreciation to Dr. Shahin Ebrahimi who has been a wise, kind and knowledgeable friend to me over the past decade. I always feel very blessed to have his valuable and reliable friendship.

Last but not the least, I would like to thank my family because of their unconditional love, respect and trust. They always encourage me to achieve my goals, realize my dreams and discover new worlds. I am truly blessed to have you. Thank you for everything you have ever done for me.

I dedicate my thesis to those who are working hard to keep peace within and among all nations despite the diversity, protect the right to be free, preserve the nature and our planet with the hope of better life for current and next generations.

Contribution of authors

- Chapter 2 to 6:

  - Narges Manouchehri: Conceptualization, methodology, software programming, validation, formal analysis, investigation, writing original draft, visualization.

  - Professor Nizar Bouguila: Supervision, reviewing and editing.

- Chapter 3 to 5:

  Dr. Wentao Fan: Reviewing.

- Chapter 2:

  Meeta Kalra: Software programming of OV-GMM.

Contents

List of Figures **xi**

List of Tables **xiii**

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

2 Online variational inference on finite multivariate Beta mixture models **6**
2.1 Introduction 7
2.2 Finite multivariate Beta mixture model 10
2.3 Batch variational learning 12
2.3.1 Prior specification 12
2.3.2 Learning algorithm 14
2.4 Online variational learning of multivariate Beta mixture models 17
2.5 Experimental results 21
2.5.1 Image segmentation in colorectal cancer 21
2.5.2 Multiclass colon tissue analysis 23
2.5.3 Digital imaging in Melanoma lesion detection and segmentation 25
2.5.4 Computer-aided detection of Malaria 26
2.6 Conclusion 28

3 A nonparametric variational learning of multivariate Beta mixture models 29
3.1 Introduction . . . 29
3.2 Model Specification . . . 31
3.2.1 Finite multivariate Beta mixture models . . . 31
3.2.2 Dirichlet process mixtures of multivariate Beta distributions . . . 33
3.3 Batch variational learning of DP mixtures of multivariate Beta distributions . . . 35
3.4 Online variational learning of DP mixtures of multivariate Beta distributions . . . 40
3.5 Results and discussion . . . 43
3.5.1 Skin lesion analysis . . . 44
3.5.2 Leukemia detection . . . 46
3.5.3 Bone tissue analysis . . . 48
3.6 Conclusion . . . 50

4 Batch and online variational learning of hierarchical Dirichlet process mixtures of multivariate Beta distributions in medical applications 51
4.1 Introduction . . . 52
4.2 Model specification . . . 54
4.2.1 Hierarchical Dirichlet process . . . 54
4.2.2 HDP mixture of multivariate Beta distributions . . . 57
4.3 Batch Variational Learning . . . 58
4.3.1 Defining prior distributions for parameters . . . 60
4.3.2 Learning algorithm . . . 61
4.4 Online variational learning of HDP mixtures of MB distributions 62
4.5 Experimental results . . . 65
4.5.1 Oral cancer diagnosis . . . 65
4.5.2 Osteosarcoma analysis . . . 68
4.5.3 Automatic white blood cell counting . . . 69
4.5.4 Discussion . . . 71
4.6 Conclusion . . . 72

5 Expectation propagation learning of finite and infinite MB mixture models 73
5.1 Introduction . . . 73

5.2 Finite multivariate Beta mixture model 75
5.3 Expectation propagation framework 77
5.4 Expectation propagation for the multivariate Beta mixture model 79
5.5 Experimental results 85
5.5.1 EEG-based sentiment analysis 85
5.5.2 Physical activity recognition 87
5.6 Conclusion 89

6 Multivariate Beta-based Hierarchical Dirichlet Process Hidden Markov Models in Medical Applications 90
6.1 Introduction 90
6.2 Model Specification 93
6.2.1 Multivariate Beta-based Hidden Markov Model 93
6.2.2 Multivariate Beta-based Hierarchical Dirichlet Process of Hidden Markov Model 95
6.3 Variational Learning 99
6.4 Experimental Results 102
6.5 Conclusion 114

7 Conclusion 115

A Proof of Equations of Chapter 2 117

B Proof of Equations of Chapter 4 122

List of References 127

List of Figures

2.1 Four examples of multivariate Beta distributions. 10
2.2 Four examples of MBMM with different components. 11
2.3 Graphical model of the finite multivariate Beta mixture. . . . 13
2.4 Sample images from colon tissues. 22
2.5 Sample images from colon dataset. 24
2.6 Sample images from skin dataset. 26
2.7 Sample images from Malaria dataset. 27

3.1 Two examples of multivariate Beta distribution. 32
3.2 Four examples of multivariate Beta mixture models with 2, 3, 4 and 5 components. 33
3.3 Graphical representation of the infinite multivariate Beta mixture model. Symbols in circles and squares denote random variables and model parameters, respectively. The conditional dependencies of the variables are represented by the arcs. The number mentioned in the plates indicates the number of repetition of the contained random variables. 37
3.4 Sample of skins. The malignant and benign cases are presented in the first and second row, respectively. 45
3.5 Confusion matrices for skin lesion analysis. 45
3.6 Examples of Leukemia dataset. The malignant and benign cases are presented in first and second row, respectively. 47
3.7 Confusion matrices for bone tissue analysis. 47
3.8 Samples of three types of bone tissues including benign, viable and necrotic tumors. 48
3.9 Confusion matrices for bone tissue analysis. 49

4.1 Examples of MB distribution and MB mixture models. 58
4.2 Examples of oral pathology dataset, benign and malignant cases in the first and second row, respectively. 67
4.3 Examples of bone pathology dataset, benign, nonviable and viable cases in the first, second and third row, respectively. . . 68
4.4 Samples of WBCs. 70

5.1 Examples of bivariate Beta distribution with three shape parameters. 76
5.2 Bivariate Beta mixture models with 2, 3, 4 and 5 components. 77
5.3 Confusion matrices for EEG analysis application. 86
5.4 Confusion matrices for human activity recognition application. 89

6.1 Multivariate Beta distribution with different shape parameters. 94
6.2 Multivariate Beta mixture models with 2, 3, 4 and 5 components. 95
6.3 Platform and sensor setup 102
6.4 Wearable sensors. 103
6.5 Different levels of activities. 103
6.6 Bar and pie chart of HAR dataset 1. 104
6.7 Oversampling results with SMOTE. 105
6.8 Examples of different ranges of features. 106
6.9 Feature distribution vs. labels. 107
6.10 Correlation matrix. 108
6.11 Bar and pie chart of HAR dataset 2. 110
6.12 Oversampling results with SMOTE. 110
6.13 Examples of different ranges of features. 111
6.14 Feature distribution vs. labels 111
6.15 Number of missing values in each feature 111
6.16 Correlation matrix 112
6.17 Results comparison. 113

List of Tables

2.1 Accuracy of image segmentation for colon cancer dataset. . . . 23
2.2 Accuracy of four models tested on multiclass colon tissue analysis. . . . 24
2.3 Accuracy of skin tumor segmentation. . . . 26
2.4 Accuracy of four models tested on Malaria dataset. . . . 28

3.1 Model performance accuracy in skin analysis. . . . 46
3.2 Model performance accuracy in Leukemia analysis. . . . 48
3.3 Model performance accuracy in bone tissue analysis. . . . 49

4.1 Model performance accuracy in oral pathology analysis. . . . 67
4.2 Model performance accuracy in bone pathology analysis. . . . 69
4.3 Model performance accuracy in WBC analysis. . . . 71

5.1 Parameter values of bivariate mixture model plots. . . . 77
5.2 Model performance accuracy in EEG-based sentiment analysis. 87
5.3 Model performance accuracy in physical activity recognition. . 88

6.1 Number of Nan in each column . . . 107
6.2 Model performance evaluation results. . . . 109
6.3 Model performance evaluation results . . . 112

List of Acronyms

ADL: Activity of Daily Living

AIC: Akaike information criterion

BIC: Bayes information criterion

BVDPMB: Batch variational learning of Dirichlet process of MB distributions

BVGMM/VR-GMM: Batch variational Gaussian mixture model

BVHDPMB: Batch variational learning of hierarchical Dirichlet process of multivariate Beta distributions

BVMBMM/VR-MBMM: Batch variational multivariate Beta mixture model

BOVW: Bag of visual words

CNS: Central nervous system

CAD/CADe: Computer aid detection

CADx: Computer-assisted diagnosis

DP: Dirichlet processes

ECG: Electroencephalography

EM: Expectation-maximization

EP: Expectation propagation

GDD: Generalized Dirichlet distribution

GMM: Gaussian mixture models

GMM-HMM: Gaussian mixture models-based Hidden Markov Models

HAR: Human activity recognition

HDP: Hierarchical Dirichlet processes

HDP-HMM: Hierarchical Dirichlet process of Hidden Markov Models

HMMs: Hidden Markov Models

IMUs: Inertial measurement units

KL: Kullback-Leibler

MCMC: Markov chain Monte Carlo

MB: Multivariate Beta distribution

MB-HMM: Multivariate Beta-based hidden Markov models

MB-HDP-HMM: Multivariate Beta-based hierarchical Dirichlet process hidden Markov models

MBMM: Multivariate Beta mixture models

MDL: Minimum description length

ML: Machine learning

ML-GMM: Maximum likelihood estimation of Gaussian mixture models

ML-MBMM: Maximum likelihood estimation of MB mixture models

MML: Minimum message length

OSCC: Oral squamous cell carcinoma

OVDPMB: Online variational learning of Dirichlet process of MB distributions

OVGMM: Online variational learning of Gaussian mixture models

OVHDPMB: Online variational learning of hierarchical Dirichlet process of multivariate Beta distributions

OVMB/OVMBMM: Online variational learning of multivariate Beta mixture model

PCA: Principal Component Analysis

SIFT: Scale-invariant feature transform

SMOTE: Synthetic Minority Over-sampling Technique

FN: False negatives

TN: True negatives

FP: False positives

TP: True positives

WBC: White blood cells

WHO: World Health Organization

Chapter 1

Introduction

1.1 Introduction and Related Work

Considering speedy developments in medical care, precision in diagnosis is one of the current demands in healthcare. This increasing need is leading to modifications to traditional practices of medicine. Such improvements result in early diagnosis, tailoring better treatment, increasing quality of patient care, and decreasing costs. Computer-assisted diagnosis [1–15] systems are among the most important improvements which have attracted lots of attention in recent years. In these methods, we use machine learning algorithms that are integrated into traditional medical equipment, analyze data and arrive at a diagnosis. These systems are considered as second opinion systems to manage diseases and improve morbidity by allowing earlier accurate detection, applying advanced diagnostics tools, tailoring better treatments, and identifying efficient curing pathways. Scientists and researchers have proposed various machine learning algorithms. However, a fraction of these solutions could be applied in practice and routine care. The roots of such limitations are some considerations and complexities in healthcare. For instance, having absolute trust in algorithms is not possible specifically when the results are not explainable in human terms [16–19]. Moreover, some tasks such as data annotation are difficult in medical cases. There are just healthcare professionals who are eligible to label data and this procedure needs lots of time and financial resources. Also, confidentiality is a major concern in medicine and this limits the amount of publicly available datasets [20].
These challenges motivated us to focus on unsupervised methods in machine

learning. Clustering and specifically mixture models have shown their capabilities in various applications. There are following issues when deploying mixture models:
1. Selecting a proper distribution to fit data: Gaussian mixture models (GMM) have been widely used in different fields [21–23]. However, despite the success of applying GMM in several domains, the assumption of Gaussianity could not be generalized to all datasets. Consequently, other distributions have been studied in recent years. In this thesis, we developed our models based on multivariate Beta distribution [24] which demonstrated considerable flexibility in fitting data [25–36]. In chapter 2, we focused on finite multivariate Beta mixture models. In chapter 3 and 4, we proposed two elegant and non-parametric frameworks, infinite and hierarchical Dirichlet process mixtures of multivariate Beta distributions, respectively. These structures empower our model by providing more flexibility. In chapter 5, we propose a novel extension of hidden Markov model, Multivariate Beta-based hierarchical Dirichlet process hidden Markov models. This new model is suggested to model sequential and temporal data.
2. Model's complexity determination: This is the second challenge and several methods have been proposed to overcome this problem. The non-parametric frameworks such as Dirichlet and hierarchical Dirichlet processes which we will discuss in chapters 3 and 4 let us produce multiple populated components while lessening the number of sparse clusters.
3. Model's parameters estimation: This is the third concern in mixture models adoption. There are several methods such as deterministic or fully Bayesian techniques to estimate the parameters. However, these methods suffer from some drawbacks such as dependency on initialization, over-fitting and convergence to local maxima in deterministic approaches. High computational time is another disadvantage in fully Bayesian inference. Variational inference and expectation propagation methods are two alternatives that are reasonable in terms of computational time and provide good accuracy also. In chapters 2, 3, 4, and 6, we learned our models with variational inference. We first considered batch setting and then extend it to online one as in real-life scenarios data arrive in an online manner. In chapter 5, we applied expectation propagation method. We evaluated our proposed models on publicly available medical datasets including images and signals. Also, we compared our models with similar alternatives in each section and report the performance of models. To assess the model performance, we used four criteria for clustering: $Accuracy = \frac{TP+TN}{\text{Total number of observations}}$, $Precision = \frac{TP}{TP+FP}$,

$Recall = \frac{TP}{TP+FN}$, $F1 - score = \frac{2 \times precision \times recall}{precision + recall}$. TP, TN, FP and FN indicate the quantity of true positives, true negatives, false positives, and false negatives, respectively. In image segmentation, we applied Jaccard Index.

1.2 Thesis Overview

This thesis is organized as follows:

- Chapter 2 presents a novel clustering method, online variational inference on finite multivariate Beta mixture models. To have simultaneous estimation of model's parameters and model complexity, variational inference technique is a proper choice. To demonstrate capability of our model, we selected challenging health-related applications including medical image analysis.

- Chapter 3 focuses on another new clustering method, nonparametric variational learning of multivariate Beta mixture models. This elegant structure provides considerable flexibility to mixture models. We measured the robustness of our proposed methods on medical images.

- Chapter 4 introduces hierarchical Dirichlet processes as another nonparametric framework and an extension of finite mixture models. We developed our unsupervised model based on multivariate Beta distribution and learned it with variational Bayesian framework. We applied our novel method to analyze medical images.

- Chapter 5 is devoted to expectation propagation learning of finite multivariate Beta mixture models. In this learning method, parameters and the complexity of the model are estimated concurrently. We conducted our research on medical signals.

- Chapter 6 introduces multivariate Beta-based hierarchical Dirichlet Process Hidden Markov Models as a powerful approach to handle observable sequential data. With this non-parametric structure and assuming that emission probabilities follow multivariate Beta mixture models, we provide more flexibility to the model. We chose variational inference to learn this model and evaluated it on medical signals.

- Chapter 7: Finally in this chapter, we conclude our work, highlight some challenges, and suggest future works.

1.3 Contributions

We summarize our main contributions as follows:

- Chapter 2: We proposed finite multivariate Beta mixture models and applied variational inference techniques with batch and online settings to learn our model. We applied this model to four medical applications including image segmentation of colorectal cancer, multi-class colon tissue analysis, digital imaging in skin lesion diagnosis, and computer aid detection of Malaria.

- Chapter 3: We extended finite multivariate Beta mixture models to the infinite case using Dirichlet Process as a non-parametric method. We used batch and online variational techniques for learning our model. To evaluate the performance of our proposed method, We applied it to three real-world medical applications, skin lesion analysis, leukemia detection, and bone tissue analysis.

- Chapter 4: We devoted this chapter to batch and online variational learning of hierarchical Dirichlet process mixtures of multivariate Beta distributions. We focused on three medical applications, oropharyngeal carcinoma diagnosis, osteosarcoma analysis, and white blood cell counting.

- Chapter 5: We applied expectation propagation inference framework to learn finite multivariate Beta mixture models and measured its performance by testing it in EEG-based sentiment analysis and human activity recognition.

- Chapter 6: We focused on analyzing sequential data with the help of Hidden Markov Models. We proposed to integrate MBMM as emission probabilities into HMM where our mixture has a nonparametric and hierarchical structure. We used variational inference to learn our proposed algorithm and evaluated our model on a health-related application, human activity recognition.

1.4 Publications and contributions of authors

This Ph.D. thesis consists of five manuscripts that represent four published journal papers and one book chapter. Each journal or book chapter is presented in a chapter of this thesis. We hereby list them:

- Chapter 2: Narges Manouchehri, Meeta Kalra, Nizar Bouguila, Online variational inference on finite multivariate Beta mixture models for medical applications, published in "IET image processing." [30].
- Chapter 3: Narges Manouchehri, Nizar Bouguila, Wentao Fan, A non-parametric variational learning of multivariate Beta mixture models in medical applications, published in "International Journal of Imaging Systems and Technology" [28].
- Chapter 4: Narges Manouchehri, Nizar Bouguila and Wentao Fan, Batch and online variational learning of hierarchical Dirichlet process mixtures of multivariate Beta distributions in medical applications, published in "Pattern Analysis and Applications" [29].
- Chapter 5: Narges Manouchehri, Nizar Bouguila and Wentao Fan, Expectation propagation learning of finite multivariate Beta mixture models and applications, published in "Neural Computing and Applications" [25]. The original published paper has a non-related medical application (forgery detection) which is removed here.
- Chapter 6: Narges Manouchehri, Nizar Bouguila, Multivariate Beta-based hierarchical Dirichlet process hidden Markov models in medical applications, published in "Hidden Markov Models and Applications" [37].

Contributions of co-authors are as follows:

- Narges Manouchehri: Conceptualization, methodology, software programming, validation, formal analysis, investigation, writing original draft, visualization.
- Professor Nizar Bouguila: Supervision, reviewing and editing.
- Meeta Kalra: Software programming of OV-GMM in chapter 2.
- Dr. Wentao Fan: Reviewing.

Chapter 2

Online variational inference on finite multivariate Beta mixture models

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

2.1 Introduction

Over the past decades, fast progress of computational power and data storage yield a great deal of complex data and machine learning methods experienced considerable development to recognize critical information from data efficiently and automatically with minimal human interaction. In order to cover the wide variety of data such as text, image and video and problem types exhibited across different domains, a diverse array of machine learning algorithms have been developed [38]. Many algorithms focus on image processing and computer vision as techniques of electronics engineering.
A critical scientific and practical goal to the majority of the algorithms is to characterize their capabilities and robustness. Supervised learning systems have been widely used over the past years. Deep learning platforms [39], [40] have been demonstrated to outperform previous supervised machine learning techniques in several fields. Convolutional Neural Networks [41] and Deep Belief Networks [42] are some examples of currently remarkable techniques in some applications such as image analysis [43], emotion detection [44], object detection [45], [46], [47], [48], synthetic aperture radar image analysis [49], [50], remote sensing [51], Internet of Things (IOT) [52], smart cities [53]. Similarly, modern medical imaging have witnessed admirable progresses and became one of the attention-grabbing domains in research and technology. Consequently, statistical modelling has been applied successfully in this domain and achieved state-of-the-art performance in image segmentation and computer-aided detection (CAD) to assist professionals in the interpretation of medical images, digital pathology and other medical datasets [54]. Due to the increasing digitization in medical image results [55] and prompt progression in artificial intelligence (AI) and machine learning (ML), various methods have been proposed [55]. However, the nature of medical data and some needs of healthcare team in making decision led to a limited success in applying the current algorithms in routine clinical cases, [16]. It should be noted that with some deep learning platforms we can achieve good results in classification tasks in various medical domains such as brain image analysis [56], pathological image analysis [57], cardiac image analysis [58], breast histology images analysis [59], blood cell analysis [60], liver tumor analysis [61]. However, they may cause some failure as they are unpredictable and unexplainable [18], [17], [19]. It should be emphasized that deep learning models need large scale labeled data for training and the publicly available datasets are limited as confidentiality is a principle rule in

healthcare. However, this is not the only issue, but medical data labeling is a great obstacle as it could be performed just by professional physicians and need sophisticated amount of budget, time and skill. It is noteworthy that the nature of medical data is heterogeneous and to arrive at a better decision, the model should have the potential and ability to deal with various types of data such as as patient history, images, videos and signals, simultaneously. These characteristics and demand, motivated us to focus on unsupervised models of machine learning as label-free approaches. Clustering methods specially finite mixture models are one of the best known methods to model heterogeneous data which includes multiple distributions [21]. The first challenging aspect which should be carefully addressed is choosing the most proper distributions that best represent the corresponding components of mixture accurately when modelling data. Gaussian mixture models (GMM) have been widely adopted in various applications [62]. However, in recent works other alternatives such as Dirichlet [63], [64], generalized Dirichlet [65], [66], [67] demonstrated considerable flexibility and high potential to describe non-Gaussian data. Hence, in our paper we focus on multivariate Beta mixture models which are developed based on a very flexible distribution which doesn't have a constant shape and is appropriate to be used to model data skewness. Furthermore, considering its bounded nature, it fits better compactly supported data. Fig. 2.1 illustrates the high potential of this distribution.

To design a clustering algorithm, the parameters estimation is a crucial step and has a significant impact on the performance of model learning. The majority of parameter estimation methods apply either deterministic or Bayesian techniques. The former one is based on classic maximum likelihood inference and optimizing the model likelihood function via expectation-maximization (EM) [68] framework. However, this method is sensitive to initialization and carry disadvantages such as over-fitting. To avoid such drawbacks, Bayesian techniques have been proposed. In this improved method, a prior knowledge is applied in a principled way and the parameter uncertainty is then marginalized by Laplace's approximation or Markov chain Monte Carlo (MCMC) simulation techniques [69], [70]. Unfortunately, we face some issues in Bayesian inference. For instance, Laplace's approximation is generally imprecise and MCMC techniques are computationally expensive. Recently, several research efforts focused on variational inference [71] as a preferable and efficient alternative technique for the learning of statistical models. Indeed, it can be expressed as an effective compromise between

deterministic and Bayesian approaches. Variational inference is based on approximating the model posterior distribution which is achieved by minimizing the Kullback-Leibler divergence between the true posterior and an approximating distribution. Another crucial issue when using mixture models is defining model structure or the best number of mixture components that describes the data perfectly without over-fitting or under-fitting. Some model selection techniques such as MML or MDL [63], [72], [73] have been considered. However, they are time-consuming since they have to evaluate a given selection criterion for several numbers of mixture components and such high computational cost limited their applications. One of the advantages of variational inference is that it automatically determines the number of mixture components as part of the Bayesian inference procedure [74], [75]. Variational learning can be performed online [76] which is mainly motivated by the fact that such algorithm allows data instances to be processed in a sequential way, which is important for large-scale data and real-time applications. This technique is significantly faster than traditional variational learning. In this paper, we propose two novel algorithms for batch and online variational learning based on multivariate Beta mixture models. We evaluate the performance of our proposed frameworks by exploring challenging medical applications and the results are compared with batch and online variational learning for Gaussian mixture models.

The structure of the rest of this paper is as follows; Section 2.2 is devoted to the description of finite multivariate Beta mixture model. Section 2.3 and 2.4 describe the batch and online variational learning algorithms, respectively. We present the experimental results in section 2.5 considering four real-world applications. Finally, we conclude in section 2.6.

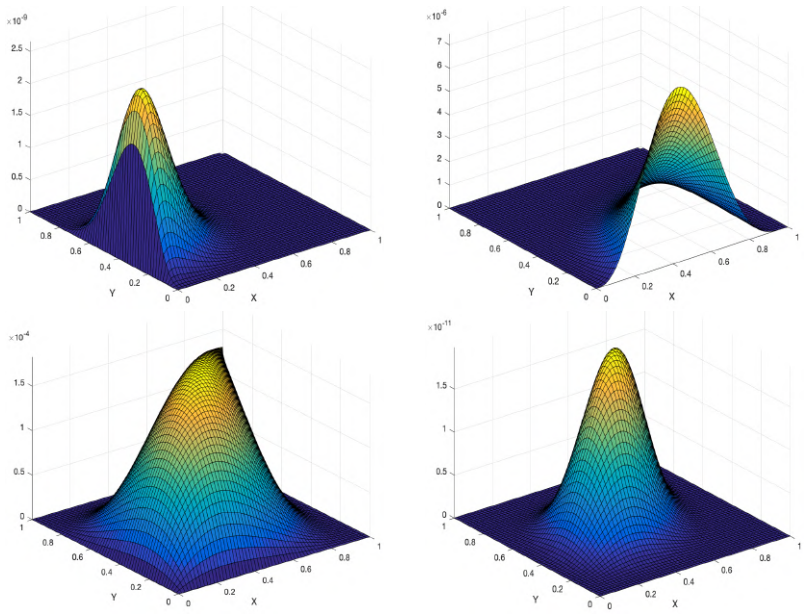


Figure 2.1: Four examples of multivariate Beta distributions.

2.2 Finite multivariate Beta mixture model

In this section, we give a brief description of finite multivariate Beta mixture models. Lets assume that an observation following a multivariate Beta (MB) distribution [77], [78] is defined by $\vec{X}_i = \left(x_{i1}, \ldots, x_{iD}\right)$ as a D-dimensional vector where all its elements are positive and less than one. $\Gamma(.)$ denotes the Gamma function. The probability density function of MB is expressed by (2.1).
$\vec{\alpha}_j = \left(\alpha_{j1}, \ldots, \alpha_{jD}\right)$ is shape parameter such that $\alpha_{jl} > 0$ for $l = 1, \ldots, D$ and $\mid \alpha_j \mid = \sum_{l=1}^{D} \alpha_{jl}$.

$$p(\vec{X}_i \mid \vec{\alpha}_j) = c \frac{\prod_{l=1}^{D} x_{il}^{\alpha_{jl}-1}}{\prod_{l=1}^{D}(1-x_{il})^{(\alpha_{jl}+1)}} \left[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1-x_{il})}\right]^{-|\alpha_j|} \tag{2.1}$$

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

Lets consider a set of N independent identically distributed vectors $\mathcal{X} =$

$\{\vec{X}_1, \ldots, \vec{X}_N\}$ which are generated from multivariate Beta mixture models and composed of M different clusters. Thus, multivariate Beta mixture model is represented by:

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{M} \pi_j p\big(\vec{X}_i \mid \vec{\alpha}_j\big) \tag{2.2}$$

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

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

Four examples of MBMM are shown in Fig. 2.2.

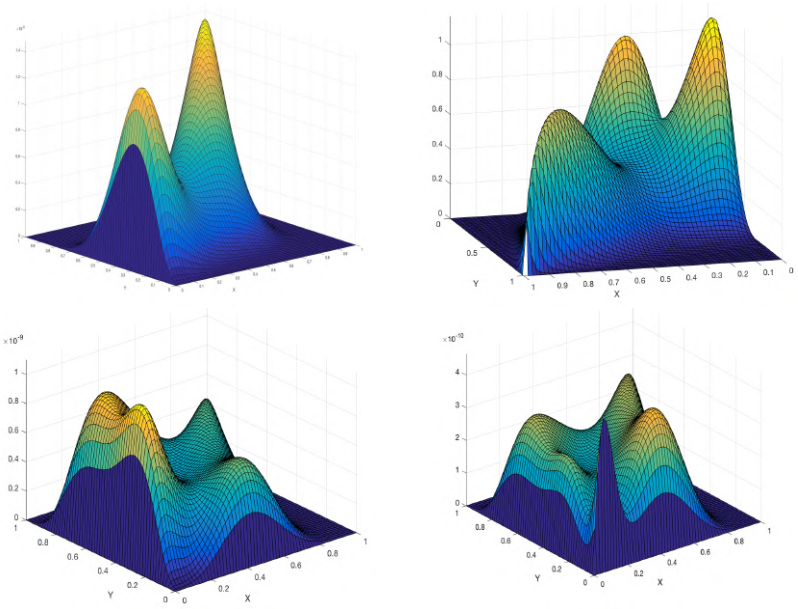


Figure 2.2: Four examples of MBMM with different components.

In mixture models, we define an auxiliary variable $\mathcal{Z}$ to allocate each sample to one of the M components. Thus, we introduce $\vec{Z}_i = \big(Z_{i1}, \ldots, Z_{iM}\big)$

where Z_{ij} is a binary random variable such that $Z_{ij} = 1$ if $\vec{X}_i$ belongs to the specific cluster j and 0, otherwise. The distribution of $\mathcal{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_N\}$ as a set of "membership vectors" is specified by (2.4) in terms of the mixing coefficients $\vec{\pi}$ [79].

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

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

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

2.3 Batch variational learning

Variational approaches have been widely applied previously to approximate posterior distributions of a variety of statistical models. In this section as the first step, we develop a batch variational inference framework for learning finite multivariate Beta mixture models (MBMM). Our main objective is to develop an optimized method which is capable enough to estimate the parameters of mixture model and determine its structure and complexity simultaneously.

2.3.1 Prior specification

A crucial challenge in the case of variational learning is placing prior distributions over parameters. To simplify this approach, we consider a conjugate prior for the $\vec{\alpha}$ parameters. Unfortunately, a conjugate prior does not exist. In this case, we adopt a Gamma prior as an approximation assuming that the parameters are statistically independent [80], [81]. So, the probability density function of α_{jl} is described by (2.6). u_{jl} and ν_{jl} are positive hyperparameters.

$$p\big(\alpha_{jl}\big) = \mathcal{G}\big(\alpha_{jl} \mid u_{jl}, \nu_{jl}\big) = \frac{\nu_{jl}^{u_{jl}}}{\Gamma\big(u_{jl}\big)} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{2.6}$$

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

$$p(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} p(\alpha_{jl}) \tag{2.7}$$

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

$$\begin{aligned} p(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}) &= p(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}) p(\mathcal{Z} \mid \vec{\pi}) p(\vec{\alpha}) \\ = \prod_{i=1}^{N} \prod_{j=1}^{M} \Bigg[& \frac{\Gamma(\mid \alpha_j \mid)}{\prod_{l=1}^{D} \Gamma(\alpha_{jl})} \times \frac{\prod_{l=1}^{D} x_{il}^{\alpha_{jl}-1}}{\prod_{l=1}^{D} (1 - x_{il})^{(\alpha_{jl}+1)}} \\ \times \Big[1 + \sum_{l=1}^{D} & \frac{x_{il}}{(1 - x_{il})} \Big]^{-|\alpha_j|} \Bigg]^{Z_{ij}} \\ \times \prod_{i=1}^{N} \Bigg[\prod_{j=1}^{M} & \pi_j^{Z_{ij}} \Bigg] \times \prod_{j=1}^{M} \prod_{l=1}^{D} \frac{\nu_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \end{aligned} \tag{2.8}$$

A graphical representation of this model is shown in Fig. 2.3. Symbols in circles denote random variables; otherwise, they denote model parameters. The conditional dependencies of the variables are represented by the arcs.

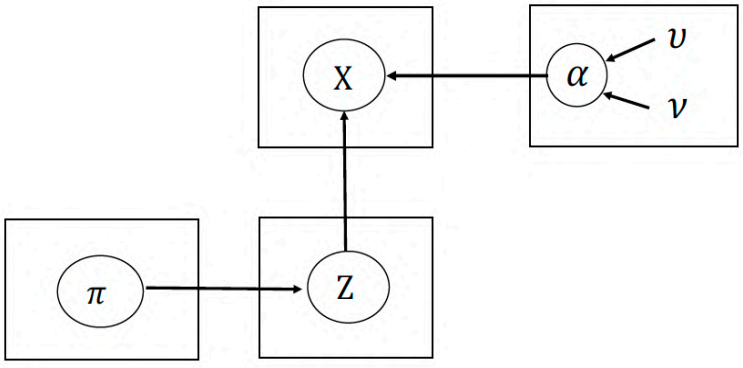


Figure 2.3: Graphical model of the finite multivariate Beta mixture.

2.3.2 Learning algorithm

In order to estimate the parameters of model and select a correct number of components, we estimate the mixing coefficient $\vec{\pi}$ by maximizing the marginal likelihood $p\big(\mathcal{X} \mid \vec{\pi}\big)$ expressed by (2.9).

$$p\big(\mathcal{X} \mid \vec{\pi}\big) = \sum_{\mathcal{Z}} \int p\big(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}\big) \mathrm{d}\vec{\alpha} \tag{2.9}$$

As the marginalization of this equation is intractable, we apply variational inference [82] to calculate the lower bound on $p\big(\mathcal{X} \mid \vec{\pi}\big)$. The variational lower bound $\mathcal{L}$ of the logarithm of the marginal likelihood $p\big(\mathcal{X} \mid \vec{\pi}\big)$ is defined by:

$$\mathcal{L}\big(Q\big) = \int Q\big(\Theta\big) \ln \left(\frac{p\big(\mathcal{X}, \Theta \mid \vec{\pi}\big)}{Q\big(\Theta\big)} \right) \mathrm{d}\Theta \tag{2.10}$$

where $\Theta = \big\{\mathcal{Z}, \vec{\alpha}, \vec{\pi}\big\}$ and $Q\big(\Theta\big)$ is an approximation to the true posterior distribution $p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)$. This approximation is determined by computation of Kullback-Leibler (KL) divergence between $Q\big(\Theta\big)$ and $p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)$ defined by (2.11).

$$KL\big(Q \mid\mid P\big) = -\int Q\big(\Theta\big) \ln \left(\frac{p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)}{Q\big(\Theta\big)} \right) \mathrm{d}\Theta \tag{2.11}$$

$$KL\big(Q \mid\mid P\big) = \ln\ p\big(\mathcal{X} \mid \vec{\pi}\big) - \mathcal{L}\big(Q\big) \tag{2.12}$$

The KL divergence is the representation of the dissimilarity between the true posterior and its approximation. As $KL\big(Q \mid\mid P\big) \geq 0$, the $KL\big(Q \mid\mid P\big)$ is zero when $Q\big(\Theta\big) = p\big(\Theta \mid \mathcal{X}\big)$. Considering above mentioned equations, it is obvious that $\mathcal{L}(Q) \leq \ln\ p\big(\mathcal{X} \mid \vec{\pi}\big)$, thus $\mathcal{L}(Q)$ is a lower bound on $\ln\ p\big(\mathcal{X} \mid \vec{\pi}\big)$. So, by maximizing the lower bound, the KL divergence is minimized and hence the true posterior distribution is approximated. Consequently, we consider a restricted and tractable family of distributions $Q\big(\Theta\big)$ which are flexible enough to properly approximate the true posterior distribution. We apply common method, namely, mean field theory to adopt factorization assumptions for restricting the form of $Q\big(\Theta\big)$. Subsequently, the posterior distribution $Q\big(\Theta\big)$ can be factorized [80] such that,

$$Q\big(\Theta\big) = Q\big(\mathcal{Z}\big) Q\big(\vec{\alpha}\big) Q\big(\vec{\pi}\big) \tag{2.13}$$

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

$$\ln Q_s^*\big(\Theta_s\big) = \langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s} \tag{2.14}$$

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

$$Q_s\big(\Theta_s\big) = \frac{exp\big\langle \ln \ p\big(\mathcal{X}, \Theta\big)\big\rangle_{i \neq s}}{\int exp\big\langle \ln \ p\big(\mathcal{X}, \Theta\big)\big\rangle_{i \neq s} \mathrm{d}\Theta} \tag{2.15}$$

$\big\langle \cdot \big\rangle_{i \neq s}$ is the expectation with respect to all the parameters other than Θ_s. The solutions for the optimal variational posteriors as derived in Appendix A are given by:

$$Q\big(\mathcal{Z}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{2.16}$$

$$Q\big(\vec{\alpha}\big) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}\big(\alpha_{jl} \mid u_{jl}^*, \nu_{jl}^*\big) \tag{2.17}$$

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M} \tilde{r}_{ij}} \tag{2.18}$$

$$\begin{aligned} \tilde{r}_{ij} = exp\Bigg\{ & \ln \pi_j + \tilde{R}_j + \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} - 1\big) \ln x_{il} \\ & - \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} + 1\big) \ln(1 - x_{il}) - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})}\Big] \Bigg\} \end{aligned} \tag{2.19}$$

where $\tilde{R}_j$ is as follows based on [83] and its calculation is presented in Appendix A.

$$u_{jl}^* = u_{jl} + \varphi_{jl}, \quad \nu_{jl}^* = \nu_{jl} - \vartheta_{jl} \tag{2.20}$$

$$\varphi_{jl} = \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \overline{\alpha}_{jl} \Bigg[\psi\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) - \psi\big(\overline{\alpha}_{jl}\big) + \sum_{s \neq l}^{D} \psi'\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) \times \overline{\alpha}_{js} \Big(\big\langle \ln \alpha_{js} \big\rangle - \ln \overline{\alpha}_{js} \Big) \Bigg] \tag{2.21}$$

$$\vartheta_{jl} = \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln x_{il} - \ln(1 - x_{il}) - \ln \Big[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})} \Big] \Bigg] \tag{2.22}$$

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

$$\big\langle Z_{ij} \big\rangle = r_{ij} \tag{2.23}$$

$$\overline{\alpha}_{jl} = \big\langle \alpha_{jl} \big\rangle = \frac{u^*_{jl}}{\nu^*_{jl}} \tag{2.24}$$

$$\big\langle \ln \alpha_{jl} \big\rangle = \psi\big(u^*_{jl}\big) - \ln \nu^*_{jl} \tag{2.25}$$

$$\Big\langle \big(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl} \big)^2 \Big\rangle = \Big[\psi\big(u^*_{jl}\big) - \ln u^*_{jl} \Big]^2 + \psi'\big(u^*_{jl}\big) \tag{2.26}$$

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{2.27}$$

In variational learning, we trace the convergence systematically by monitoring the variational lower bound during the re-estimation step. Indeed, at each step of the iterative updating procedure, the value of $\mathcal{L}(Q)$ should not rise. Thus, we terminate optimization when the lower bound increases more than a threshold compared to previous estimated value. The lower bound in (2.10) is evaluated as follows which is explained in details in Appendix A:

$$\mathcal{L}\big(Q\big) = \sum_{\mathcal{Z}} \int Q\big(\mathcal{Z}, \vec{\alpha}\big) \ln \Bigg(\frac{p\big(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}\big)}{Q\big(\mathcal{Z}, \vec{\alpha}\big)} \Bigg) \mathrm{d}\vec{\alpha} \tag{2.28}$$

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

Algorithm 1 Batch variational framework for MBMM.

1. Choose a large initial number of components M.
2. Initialize values for u_{jd} and v_{jd}.
3. Initialize the value of r_{ij} using K-Means algorithm.
4. repeat
5. The variational E-step:
6. Estimate the expected values by equations (2.23) to (2.27).
7. The variational M-step:
8. Update $Q(\mathcal{Z})$ and $Q(\vec{\alpha})$ by estimating r_{tj} from (2.16) and (2.17).
9. until Convergence criterion is reached.

2.4 Online variational learning of multivariate Beta mixture models

In this subsection, we extend the classic variational inference approach [81] to online settings for learning multivariate Beta mixture model by adopting the framework proposed in [76] as in real-world, observations arrive in an online manner. Thus, we assume that a specific amount of data are observed defined by t, such that their corresponding lower bound is defined by [84]:

$$\mathcal{L}^{(t)}(Q) = \frac{N}{t} \sum_{i=1}^{t} \int Q(\alpha) \mathrm{d}\alpha \sum_{\vec{Z}_i} Q(\vec{Z}_i) \ln \left[\frac{p(\vec{X}_i, \vec{Z}_i \mid \alpha)}{Q(\vec{Z}_i)} \right] + \int Q(\alpha) \ln \left[\frac{p(\alpha)}{Q(\alpha)} \right] \mathrm{d}\alpha \tag{2.29}$$

In this method, the current variational lower bound expressed by (2.29) is maximized consecutively. To explain more in detail, lets consider a set of observations $\{\vec{X}_1, \ldots, \vec{X}_{(t-1)}\}$. Then, a new observation $\vec{X}_t$ arrives and we maximize and update the current lower bound $\mathcal{L}^{(t)}(Q)$ corresponding to $Q(\vec{Z}_t)$, while $Q(\vec{\alpha})$ and π_j is set to $Q^{t-1}(\vec{\alpha})$ and π_j^{t-1}, respectively. Thus, the variational solution to $Q(\vec{Z}_t)$ is as follows:

$$Q\left(\vec{Z}_t\right) = \prod_{j=1}^{M} r_{tj}^{Z_{tj}} \tag{2.30}$$

$$r_{ij} = \frac{\tilde{r}_{tj}}{\sum_{j=1}^{M} \tilde{r}_{tj}} \tag{2.31}$$

$$\begin{aligned} \tilde{r}_{tj} = exp\Bigg\{ & \ln \pi_j^{(t-1)} + \tilde{R}_j + \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} - 1\right) \ln x_{tl} \\ & - \sum_{l=1}^{D} \left(\overline{\alpha}_{jl} + 1\right) \ln(1 - x_{tl}) - \mid \overline{\alpha}_j \mid \ln \left[1 + \sum_{l=1}^{D} \frac{x_{tl}}{(1 - x_{tl})}\right] \Bigg\} \end{aligned} \tag{2.32}$$

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

$$Q^{(t)}\left(\vec{\alpha}\right) = \prod_{j=1}^{M} \prod_{d=1}^{D} \mathcal{G}\left(\alpha_{jd}^{(t)} \mid u_{jd}^{*(t)}, v_{jd}^{*(t)}\right) \tag{2.33}$$

$$u_{jd}^{*(t)} = u_{jd}^{*(t-1)} + \rho_t \Delta u_{jd}^{*(t)} \tag{2.34}$$

$$v_{jd}^{*(t)} = v_{jd}^{*(t-1)} + \rho_t \Delta v_{jd}^{*(t)} \tag{2.35}$$

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

$$\pi_j^{(t)} = \pi_j^{(t-1)} + \rho_t \Delta\pi_j^{(t)} \tag{2.36}$$

where ρ_t denotes the learning rate [85] described by (2.37) with two constraints, $\epsilon \in (0.5, 1]$ and $\eta \geq 0$.

$$\rho_t = (\eta_0 + t)^{-\epsilon} \tag{2.37}$$

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

$$\begin{aligned}
&\Delta u_{jl}^{*(t)} = u_{jl}^{*(t)} - u_{jl}^{*(t-1)} = \\
&u_{jl} - +Nr_{tj}\overline{\alpha}_{jl}\Bigg[\psi\left(\sum_{l=1}^{D}\overline{\alpha}_{jl}\right) - \psi\left(\overline{\alpha}_{jl}\right) \\
&+ \sum_{d\neq s}^{D}\psi'\left(\sum_{l=1}^{D}\overline{\alpha}_{jl}\right) \times \overline{\alpha}_{js}\left(\langle \ln \alpha_{js}\rangle - \ln \overline{\alpha}_{js}\right)\Bigg]
\end{aligned} \tag{2.38}$$

$$\begin{aligned}
&\Delta v_{jl}^{*(t)} = v_{jl}^{*(t)} - v_{jl}^{*(t-1)} = v_{jl} \\
&- Nr_{tj}\left[\ln x_{td} - \ln(1 - x_{tl}) - \ln\left[1 + \sum_{l=1}^{D}\frac{x_{tl}}{(1 - x_{tl})}\right]\right]
\end{aligned} \tag{2.39}$$

$$\Delta\pi_j^{(t)} = \pi_j^{(t)} - \pi_j^{(t-1)} = \left(\frac{N}{t}\right) r_{tj} - \pi_j^{(t-1)} \tag{2.40}$$

where $\left\langle \ln \alpha_{jd}\right\rangle$ and $\left\langle \left(\ln \alpha_{jd} - \ln \overline{\alpha}_{jd}\right)^2\right\rangle$ are similar to (2.25) and (2.26), respectively. When a new data point arrives, an additional distribution is added to the lower bound.
There are two constraints expressed by (2.41) that ensure the convergence

of lower bound as the online learning framework can be considered as a stochastic approximation:

$$\sum_{t=1}^{\infty} \rho_t = \infty, \qquad \sum_{t=1}^{\infty} {\rho_t}^2 < \infty \tag{2.41}$$

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

Algorithm 2 Online variational framework for MBMM.

1. Choose a large initial number of components M.
2. Initialize values for u_{jd} and v_{jd}.
3. Initialize the value of r_{ij} using K-Means algorithm.
4. For $t = 1$ to N
5. repeat
6. The variational E-step:
7. Estimate the expected values.
8. Calculate learning rate by (2.37).
9. Compute $\Delta u_{jd}^{*(t)}$, $\Delta v_{jd}^{*(t)}$ and $\Delta \pi_j^{(t)}$ by (2.38) to (2.40).
10. The variational M-step:
11. Update the variational solutions to update $Q(\mathcal{Z}_t)$.
12. Update the variational solutions to update $Q(\vec{\alpha})$.
13. until Convergence criterion is reached.

2.5 Experimental results

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

2.5.1 Image segmentation in colorectal cancer

According to World Health Organization (WHO) reports, cancer is the second leading cause of death globally, taking the life of 1 in 6 people, accounting for an estimated 9.6 million deaths in 2018 [86]. Colorectal cancer with 1.80 million cases, has the third place in the ranking of most common cancers and secondly ranked in most typical causes of cancer death with 862,000 deaths.

Early detection and treatment has a great impact on reducing cancer mortality. By early identification and avoiding delays in care, the patient is more likely to survive by responding effectively to treatments. This goal is achieved by awareness, accessing clinical evaluation, diagnosis and having access to treatment [86]. One of the valuable solutions to avoid late stages detection is screening which aims to find individuals with abnormalities, pre-cancer and not developed symptoms. As one of the main steps of screening, tissue or cell samples can be taken from intestine or stomach for determining causes of abnormalities or presence and effects of cancer. Hence, histopathology analysis has a significant role and poses critical challenge as biological tissues have various structures and precise tumor segmentation, accurate pattern detection is a tough task for humans. In recent years, since tissue specimens were digitized, automated analysis of histopathology slides [87] has become a key requirement to asses quantitative morphology, cancer aggressiveness grading and reliable differentiation of various tumor types which is reflected by the formation and architecture of glands. Subsequently, machine learning techniques have demonstrated superior performance over conventional methods [88]. Here we focus on two applications related to colorectal cancers. First, image segmentation of a publicly available collection of microscopy images of colon cancer cells from Broad Bioimage Benchmark Collection (BBBC018v1) [89], [90]. The image set consists of 56 fields of view (4 from each of 14 samples). Because there are three channels, there are 168 image files. The samples were stained with Hoechst 33342, pH3, and phalloidin. Hoechst 33342 is a DNA stain that labels the nucleus. Phospho-histone H3 indicates mitosis. Phalloidin labels actin, which is present in the cytoplasm. This image set is accompanied by a set of ground truth data to test automated image analysis against them. The ground truth set consists of outlines of nuclei and cells. In Fig. 2.4, some examples of tissues and nucleous with their corresponding ground truth are illustrated.

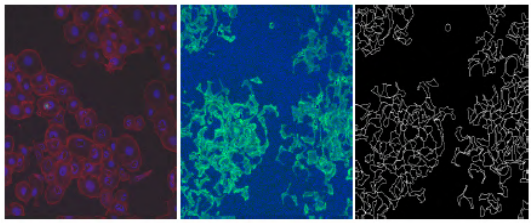


Figure 2.4: Sample images from colon tissues.

The results of validating our proposed frameworks based on Jaccard similarity index are presented in Table 2.1 which prove that our model outperforms the three other alternatives.

Table 2.1: Accuracy of image segmentation for colon cancer dataset.

Method	Accuracy
OVMBMM	95.77
OVGMM	87.41
BVMBMM	90.31
BVGMM	84.33

2.5.2 Multiclass colon tissue analysis

The second application is multiclass tissue clustering problem and categorization of a collection of textures in histological images of human colorectal cancer. The term texture refers to specific properties, pattern and structure of image regions. In medical image analysis, texture analyzing methods are applied to classify tissue types. Human solid tumours are complex structures that typically several distinct tissue types are integrated in tumors consisting of non-malignant tissues, necrotic regions, tumour stroma, immune cell infiltration and islets of remaining. Moreover, tumour progression over time leads to changes in the architecture of tissue. In the digital pathology, automatic recognition of different tissue types assists to estimate the tumour/stroma ratio on histological samples and can provide quantitative and high-throughput analysis of the tumour tissue. In this paper to assess the performance, we evaluate our models by a collection of textures in colorectal cancer histology [91], [92] which is publicly available. It includes 5,000 histological images of human colorectal cancer consisting of eight different types of tissue. In Fig. 2.5, some samples of eight tissue classes are shown which enhance a variety of illumination, stain intensity and tissue textures. These classes are tumour epithelium, simple stroma that is homogeneous composition including tumour or extra-tumoural stroma, smooth muscle, single tumour or immune cells and/or single immune cell, complex stroma containing single tumour cells and/or few immune cells, immune cells including immune-cell conglomerates and sub-mucosal lymphoid follicles, debris including necrosis, hemorrhage and mucus, normal mucosal glands, adipose tissue and background without any tissue. As an important step, we extracted the feature of each image

using one of the most popular techniques, namely, scale-invariant feature transform (SIFT) [93] and bag of visual words (BOVW). The general idea of this method is to represent an image as a set of features which include key points and descriptors. The keypoints of each image are invariant to geometrical transformation and illumination and descriptors are the description of these points which both are extracted by SIFT. Consequently, we construct vocabularies with key points and descriptors to represent each image as a frequency histogram of features which could be applied in image categorization to find images with similar pattern which could be differentiated by histopathological evaluation and the tissue composition could be quantified. The outputs of testing the performance of our algorithms are illustrated in Table 2.2 which obviously shows the superior performance of OVMBMM.

Table 2.2: Accuracy of four models tested on multiclass colon tissue analysis.

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

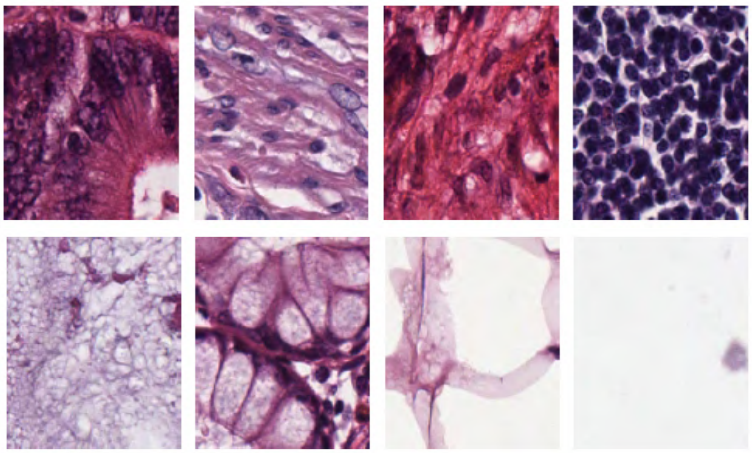


Figure 2.5: Sample images from colon dataset.

2.5.3 Digital imaging in Melanoma lesion detection and segmentation

As stated by WHO, 1.04 million cases of skin cancer were reported in 2018 and it was ranked as the 5th common cancer [86]. The major cause of death from skin cancer is malignant melanoma which is caused by the abnormal multiplication of cells. However, it is far less prevalent than non-melanoma skin cancers. This type of cancer is primarily diagnosed visually. After initial clinical screening and dermoscopic analysis, a biopsy and histopathological sample is analyzed. Digital imaging can help to recognize and treat in its earliest stages which lead to reduce melanoma mortality as it is readily curable. Automated diagnosis and digital images of skin lesions can aid in the diagnosis of melanoma through teledermatology. The standard quality of skin lesion imaging has a great impact on early detection and results in improvement of the efficiency, effectiveness, and accuracy of melanoma diagnosis. Nevertheless, unprofessional screening results in unnecessary biopsies and excisions of benign skin lesions. However, it is difficult to distinguish early-stage melanoma from benign skin lesions with similar structure which may lead to missing positive cases, useless clinical advanced examinations and misclassifying the benign and malignant melanoma. Thus, the expertise of the examiner and clinical setting have significant role. Evolution of digital imaging in skin lesion diagnosis permits the early detection of atypical lesions. Therefore, unnecessary biopsies of benign tumors are decreased or avoided. Recent enhancements in computer vision, machine learning algorithms and digital dermoscopic techniques can assist in image segmentation and retrieval, facilitate follow up and reduce unbiased diagnosis and misclassification rate therefore. These admirable advantages led to gaining the attention of researchers and increasing the focus towards computer aided systems in the last few decades. To evaluate automated Melanoma region segmentation using dermoscopic images, we tested our models, on a public dataset of ISIC [94], containing 23,906 images of skin lesions with their corresponding ground truths. In Fig. 2.6, some samples of this dataset and their ground truth are illustrated. Similar to previous experiments, we compared four models. The outputs are presented in Table 2.3 based on Jaccard similarity index. As it is shown, OVMBMM is more accurate than the other algorithms.

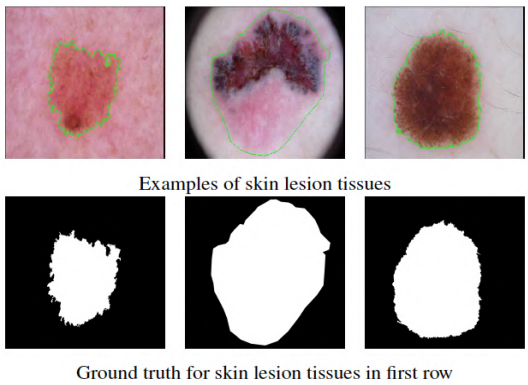


Figure 2.6: Sample images from skin dataset.

Table 2.3: Accuracy of skin tumor segmentation.

Method	Accuracy
OVMBMM	94.29
OVGMM	87.53
BVMBMM	92.09
BVGMM	82.77

2.5.4 Computer-aided detection of Malaria

Malaria is a serious infectious disease caused by a blood parasite which is injected into the human body by female Anopheles mosquito. Considering the statistics announced by WHO, 219 million Malaria cases and 435 000 Malaria deaths were globally reported in 2017 [95]. To manage and monitor this disease efficiently, it is crucial to diagnose it promptly and accurately as misdiagnosis can lead to significant morbidity and mortality. Therefore, with the help of parasitological and clinical microscopy which is considered as the mainstay of parasite-based diagnosis, the infection could be identified and confirmed precisely. The microscopy examination of Malaria, as the most prevalent and commonly practiced method, involves visual examination blood smears to test for the presence or absence of parasite in the blood and quantification of parasitemia, specie identification and life cycle classification. However, we should bear in mind that acceptable microscopy service with consistently accurate results is time consuming and costly and depends on

the qualification of experts and load of samples. WHO reported that more than 208 million patients were tested by microscopic examination in 2017. Such massive number of ongoing examinations indicates the significance of process automation in analysis of samples. In order to overcome the issues such as error-prone and timely procedure, computer aid detection (CAD) and mathematical morphology are applied as effective tools for computer aided Malaria detection. These techniques are widely used for image processing purposes and employed successfully in biomedical image analysis. However, computer vision techniques for diagnosis, recognition and differentiation between non-parasitic and infected samples represent a relatively new domain of research. In our work, we applied our models on a dataset provided by NIH including thin blood smear slide images from the Malaria Screener research activity [96]. The dataset contains a total of 27,558 cell images with equal instances of parasitized and uninfected cells. A few examples of this dataset are illustrated in Fig. 2.7 including parasitized and uninfected blood smear samples. In this experiment, the features are extracted by BOVW and SIFT. Finally to evaluate the performance of our method, we compared the results of four models which are illustrated in Table 2.4 indicating that OVMBMM has more accurate outputs. It is noteworthy to mention that these results clarify that online variational learning is a robust method as physicians are analyzing large amount of pathological samples.

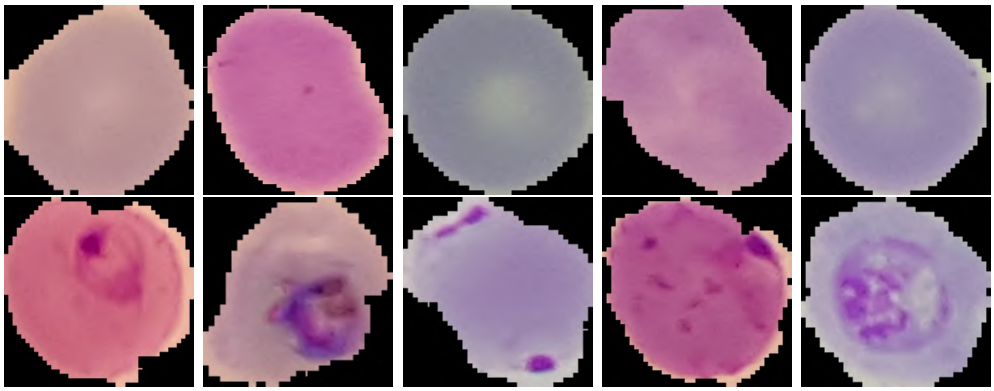


Figure 2.7: Sample images from Malaria dataset.

Table 2.4: Accuracy of four models tested on Malaria dataset.

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

2.6 Conclusion

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

Chapter 3

A nonparametric variational learning of multivariate Beta mixture models

Clustering as an essential technique has matured into a capable solution to address the gap between growing availability of data and deriving the knowledge from them. In this paper, we propose a novel clustering method "variational learning of infinite multivariate Beta mixture models". The motivation behind proposing this technique is flexibility of mixture models to fit the data. This approach has the capability to infer the model complexity and estimate model parameters from the observed data automatically. Moreover, as a label-free method, it could also address the problem of high costs of medical data labeling which can be undertaken just by medical experts. The performance of the model is evaluated on real medical applications and compared to other similar alternatives. We demonstrate the ability of our proposed method to outperform widely used methods in the field as it has been shown in experimental results.

3.1 Introduction

Analytical algorithms have been widely used to extract hidden structures of data in past couple of decades. Among all the attention grabbing approaches, clustering has been the topic of various researches. The main concept in the unsupervised algorithms is to automatically cluster a dataset into various groups. Finite mixture models have been applied for clustering

as they demonstrated considerable flexibility to express data [97], [22] and the integration of prior knowledge about the data is easily possible. In finite mixture models, we need to select a proper distribution to adequately fit data. Gaussian mixture models (GMM) have been broadly adopted in various real-world applications so far. However, the nature of all data is not generally Gaussian. In recent years, other distributions such as Dirichlet [98], [99], Gamma and Beta [100], [101], [102], [103], inverted Dirichlet [104], generalized Dirichlet [105], generalized inverted Dirichlet [106], Beta-Liouville [107], [108] and inverted Beta-Liouville [109] have been chosen to tackle the limitations of non-Gaussianity. Moreover, defining model complexity is another task in applying mixture models that has been handled by some selection criteria like Akaike information criterion (AIC) [110], Bayes information criterion (BIC) [111], Minimum description length (MDL) and minimum message length (MML) [112], [113]. Due to the drawback of these methods such as being time consuming, other alternatives have been introduced. Dirichlet Process (DP) as a non-parametric method and an extension to finite mixture model is one of the solutions to mitigate this problem [114]. Similarly, we are concerned about parameter estimation of mixture models. Several frequentist and Bayesian techniques have been extensively proposed such as expectation maximization [115] and Markov Chain Monte Carlo [116], [117], [118], [119]. The deterministic methods have some disadvantages such as sensitivity to initialization and convergence to local maxima/saddle points [120]. Bayesian techniques have complex computational processing [121]. To overcome these barriers, variational inference was introduced as an alternative approach [122], [123], [124] which approximates the model posterior distribution by minimizing the Kullback-Leibler divergence between the true posterior and an approximating distribution. In this paper, we propose a novel method based on variational learning of Dirichlet process mixtures of multivariate Beta distributions [32] to compute model's complexity and estimate model parameters simultaneously. To evaluate the goodness of model performance, we applied it on three real world challenging medical applications, namely, skin lesion analysis, leukemia detection and bone tissue analysis. Our experimental results demonstrates the practicality of our proposed model in healthcare as it doesn't need labeling various types of data. Moreover, interpretation of complex biomedical data in variable health scenarios is a challenging task even for qualified physicians. We show how our automated algorithm could help in analyzing heterogeneous data. We can summarize our four folded contributions as follows:

1. We study the behaviour of multivariate Beta distribution [125], [126] to construct a DP mixture based on it. Our motivation is its high flexibility.

2. To learn our model, we apply batch variational inference methodology to estimate model parameters and determine model complexity automatically at the same time.

3. Then, the batch variational learning will be extended to online setting. This solution is an alternative to tackle the problem of dealing with sequential data particularly in real-world applications.

4. We select challenging medical applications which to the best of our knowledge haven't been widely considered for studying similar algorithms. The robustness of our proposed model will be analyzed comparing it to other similar models.

The remainder of this work is organized as follows: In section 3.2, we review finite multivariate Beta mixture models and then extend it to the infinite case. In section 3.3, we discuss batch variational framework of infinite multivariate Beta mixture models followed by Section 3.4 which is dedicated to the extension to online setting. In Section 3.5, we present the experimental results. Finally, we conclude in section 3.6.

3.2 Model Specification

In this section, first we present multivariate Beta (MB) distribution. Then, we express how multivariate Beta mixture models are constructed. Afterward, we extend finite mixture models to the infinite form.

3.2.1 Finite multivariate Beta mixture models

Let's assume a dataset $\mathcal{Y} = \{\vec{Y}_1, \ldots, \vec{Y}_N\}$ containing N observations which are independent and identically distributed. $\vec{Y}_i = \big(y_{i1}, \ldots, y_{iD}\big)$ as a D-dimensional vector represents an observation raised from a MB distribution with following probability density function such that $0 < y_{id} < 1$ and $\Gamma < . >$ indicates the Gamma function:

$$p(\vec{Y_i} \mid \alpha_j) = \frac{c \prod_{d=1}^{D} y_{id}^{\alpha_{jd}-1}}{\prod_{d=1}^{D}(1-y_{id})^{(\alpha_{jd}+1)}} \Big[1 + \sum_{d=1}^{D} \frac{y_{id}}{(1-y_{id})}\Big]^{-|\alpha_j|}, \quad c = \frac{\Gamma(\mid \alpha_j \mid)}{\prod_{d=1}^{D} \Gamma(\alpha_{jd})} \tag{3.1}$$

$\vec{\alpha_j} = \big(\alpha_{j1}, \ldots, \alpha_{jD}\big)$ represents the shape parameter where $\alpha_{jd} > 0$ for $d = 1, \ldots, D$ and $\mid \alpha_j \mid = \sum_{d=1}^{D} \alpha_{jd}$. To construct a finite mixture based on multivariate Beta distributions [22] which includes M components, we assume that all vectors of $\mathcal{Y}$ follow common probability density function given by:

$$p\big(\vec{Y_i} \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{M} \pi_j p\big(\vec{Y_i} \mid \vec{\alpha}_j\big) \tag{3.2}$$

$\vec{\alpha} = \big(\vec{\alpha}_1, \ldots, \vec{\alpha}_M\big)$ indicates set of shape parameters for all M clusters. $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ represents the components weights and π_j is mixing coefficient of component j subjected to two constraints $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j >= 0$. Thus, the likelihood function for N data points is expressed by:

$$p\big(\mathcal{Y} \mid \vec{\pi}, \vec{\alpha}\big) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j p\big(\vec{Y_i} \mid \vec{\alpha}_j\big) \right] \tag{3.3}$$

It is noteworthy to mention that our motivation to choose multivariate Beta distribution as our basic distribution is its substantial flexibility and high potential to properly fit data. Figure 3.1 and Figure 3.2 illustrate two examples of MB distribution and four examples of its mixture models.

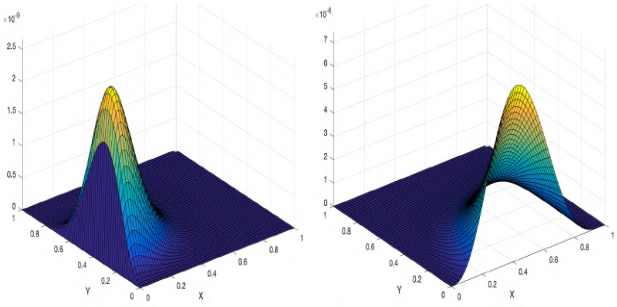


Figure 3.1: Two examples of multivariate Beta distribution.

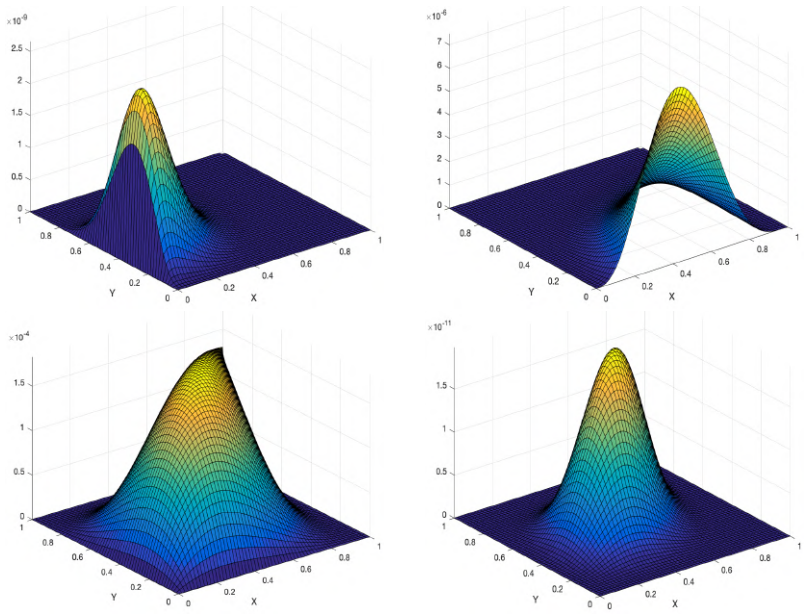


Figure 3.2: Four examples of multivariate Beta mixture models with 2, 3, 4 and 5 components.

3.2.2 Dirichlet process mixtures of multivariate Beta distributions

One major challenge in mixture modelling is the determination of model complexity and defining number of mixture components that best describes the data without over or under-fitting. To bypass this obstacle, we extend the finite mixture model to the infinite case by adopting a framework of Dirichlet process (DP) mixture model. In this case, we assume that the number of clusters, M is infinite. To construct the model, we adopt a stick-breaking representation of DP [127], [128]. Lets assume that DP is defined by G with H and β as base distribution and scaling parameter, respectively. The stick-breaking construction of $G \sim DP(\beta, H)$ is expressed by equation (3.4). δ_{Ω_j} indicates the Dirac delta measure centered at Ω_j and β is a real number. To find π_j, we apply a recursively breaking procedure to a unit length stick and

break it into an infinite number of pieces such that $\sum_{j=1}^{\infty} \pi_j = 1$.

$$\lambda \sim Beta(1, \beta), \quad \Omega_j \sim H, \quad \pi_j = \lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s), \quad G = \sum_{j=1}^{\infty} \pi_j \delta_{\Omega_j} \tag{3.4}$$

Given an observed dataset $\mathcal{Y} = \{\vec{Y}_1, \ldots, \vec{Y}_N\}$ generated from a MB mixture model with a countably infinite number of components, for each observation we have:

$$p\big(\vec{Y} \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{\infty} \pi_j \prod_{d=1}^{D} p\big(y_{id} \mid \alpha_{jd}\big) \tag{3.5}$$

Afterward, we define an auxiliary and binary latent variable $\mathcal{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_N\}$ to allocate each sample to one of the components. Over each observation $\vec{Y}_i$, we place a vector $\vec{Z}_i = \big(Z_{i1}, Z_{i2}, \ldots\big)$ such that $Z_{ij} \in \{0, 1\}$, $\sum_{j=1}^{\infty} Z_{ij} = 1$ and $Z_{ij} = 1$ if $\vec{Y}_i$ is assigned to the specific cluster j and 0, otherwise. $\mathcal{Z}$ as a set of "membership vectors" is formulated by equation (3.6) in terms of $\vec{\pi}$ [129].

$$p\big(\mathcal{Z} \mid \vec{\pi}\big) = \prod_{i=1}^{N} \prod_{j=1}^{\infty} \pi_j^{Z_{ij}} \tag{3.6}$$

According to the stick-breaking construction of DP, π_j is a function of λ_j and $p\big(\mathcal{Z} \mid \vec{\pi}\big)$ can be expressed as [130]:

$$p\big(\mathcal{Z} \mid \vec{\lambda}\big) = \prod_{i=1}^{N} \prod_{j=1}^{\infty} \left[\lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s)\right]^{Z_{ij}} \tag{3.7}$$

The complete likelihood function of the infinite multivariate Beta mixture model with latent variables as the conditional distribution of $\mathcal{Y}$ given the class labels $\mathcal{Z}$ is defined as:

$$p\big(\mathcal{Y} \mid \mathcal{Z}, \vec{\alpha}\big) = \prod_{i=1}^{N} \prod_{j=1}^{\infty} \left(\prod_{d=1}^{D} p\Big(y_{id} \mid \alpha_{jd}\Big)\right)^{Z_{ij}} \tag{3.8}$$

3.3 Batch variational learning of DP mixtures of multivariate Beta distributions

As we explained in introduction, one of our concerns when applying mixture models is parameters estimation. The major objective in this section is to develop an optimized and capable method for estimating our model parameters which are defined by $\Theta = \{\mathcal{Z}, \vec{\alpha}, \vec{\lambda}, \vec{\beta}\}$. To achieve this objective, we use variational inference. Based on the methodology of this technique, we find an approximation $Q(\Theta)$ for the true posterior distribution $p(\Theta \mid \mathcal{Y})$ [131], [123], [132]. The dissimilarity between $Q(\Theta)$ and $p(\Theta \mid \mathcal{Y})$ is computed by Kullback-Leibler (KL) divergence defined by equation (3.9) where $\mathcal{L}$ indicates lower bound:

$$KL\big(Q \parallel P\big) = -\int Q\big(\Theta\big) \ln \left(\frac{p\big(\Theta \mid \mathcal{Y}\big)}{Q\big(\Theta\big)}\right) \mathrm{d}\Theta = \ln\ p\big(\mathcal{Y} \mid \vec{\pi}\big) - \mathcal{L}\big(Q\big) \quad (3.9)$$

$$\mathcal{L}\big(Q\big) = \int Q\big(\Theta\big) \ln \Big(p\big(\mathcal{Y}, \Theta \mid \vec{\pi}\big)/Q\big(\Theta\big)\Big) \mathrm{d}\Theta \quad (3.10)$$

As KL divergence is subjected to the constraint of $KL(Q \parallel P) \geq 0$, this measure reaches to zero when $Q(\Theta) = p(\Theta \mid \mathcal{X})$ and $\mathcal{L}(Q) \leq \ln p(\mathcal{X})$. Thus, we can conclude that $\mathcal{L}(Q)$ is a lower bound to $\ln p(\mathcal{X})$. Considering the intractability of true posterior distribution, its direct calculation is computationally complex. To overcome this problem, a restricted family of $Q(\Theta)$ that can be computed is considered [123]. Borrowing the idea from mean field theory [132], we factorize $Q(\Theta)$ into disjoint distributions which are tractable such that $Q(\Theta) = \prod_i Q_i(\Theta_i)$. By adopting variational learning with respect to each $Q_i(\Theta_i)$, the lower bound $\mathcal{L}(Q)$ is maximized. For Q_s as a specific factor, we fix $\{\Theta_i\}_{i \neq s}$ and maximize $\mathcal{L}(Q)$ with respect to all possible forms for the distribution $Q_s(\Theta_s)$ [132] and the optimal solution for a specific factor, $Q_s\big(\Theta_s\big)$, is given by equation (3.11). $\langle \cdot \rangle_{i \neq s}$ indicates an expectation with respect to all the distributions $Q_i(\Theta_i)$ except for $i = s$ [132].

$$Q_s\left(\Theta_s\right) = \frac{\exp\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s}}{\int \exp\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s} d\Theta} \quad (3.11)$$

To learn infinite mixture models by variational method, we should exploit the bound by defining a truncation of the stick-breaking representation [124],

[133]. Here, this truncation level is shown by M such that $\lambda_M = 1$ and $\pi_j = 0$ when $j > M$ leading to $\sum_{j=1}^{M} \pi_j = 1$.
To develop variational learning, defining prior distributions over parameters is one of the main concerns. Based on equation (3.4), a Beta distribution is placed on $\vec{\lambda}$ and $\vec{\beta} = (\vec{\beta}_1, \vec{\beta}_2, \dots)$ are the hyperparameters of $\vec{\lambda}$.

$$p\big(\vec{\lambda} \mid \vec{\beta}\big) = \prod_{j=1}^{\infty} Beta(1, \beta_j) = \prod_{j=1}^{\infty} \beta_j (1 - \lambda_j)^{\beta_j - 1} \tag{3.12}$$

As suggested in [127], the Gamma distribution is introduced as conjugate prior to the stick lengths over β as defined in equation (3.13) where $\vec{a} = (a_1, a_2, \dots)$ and $\vec{b} = (b_1, b_2, \dots)$ are its positive hyperparameters.

$$p(\vec{\beta}) = \mathcal{G}\big(\vec{\beta} \mid \vec{a}, \vec{b}\big) = \prod_{j=1}^{\infty} \frac{b_j^{a_j}}{\Gamma(a_j)} \beta_j^{a_j - 1} e^{-b_j \beta_j} \tag{3.13}$$

Next, the Gamma distribution as the prior distributions for α_{jd} [134] is adopted where its hyperparameters $\vec{u}_{jd}$ and $\vec{\nu}_{jd}$ are positive:

$$p\big(\alpha_{jd}\big) = \mathcal{G}\big(\alpha_{jd} \mid u_{jd}, \nu_{jd}\big) = \frac{\nu_{jd}^{u_{jd}}}{\Gamma\big(u_{jd}\big)} \alpha_{jd}^{u_{jd} - 1} e^{-\nu_{jd} \alpha_{jd}} \tag{3.14}$$

We assume all the parameters to be statistically independent:

$$p\big(\vec{\alpha}\big) = \prod_{j=1}^{\infty} \prod_{d=1}^{D} p\big(\alpha_{jd}\big) = \frac{v_{jd}^{u_{jd}}}{\Gamma(u_{jd})} \alpha_{jd}^{u_{jd} - 1} e^{-v_{jd} \alpha_{jd}} \tag{3.15}$$

Thus, the joint distribution of all the random variables is given by (3.16):

$$p\big(\mathcal{Y}, \Theta\big) = p\big(\mathcal{Y} \mid \mathcal{Z}, \vec{\alpha}\big) p\big(\mathcal{Z} \mid \vec{\lambda}\big) p\big(\vec{\lambda} \mid \vec{\beta}\big) p\big(\vec{\beta}\big) p\big(\vec{\alpha}\big) \tag{3.16}$$

Figure 3.3 is a graphical representation of this model to illustrate the dependencies between all the variables.

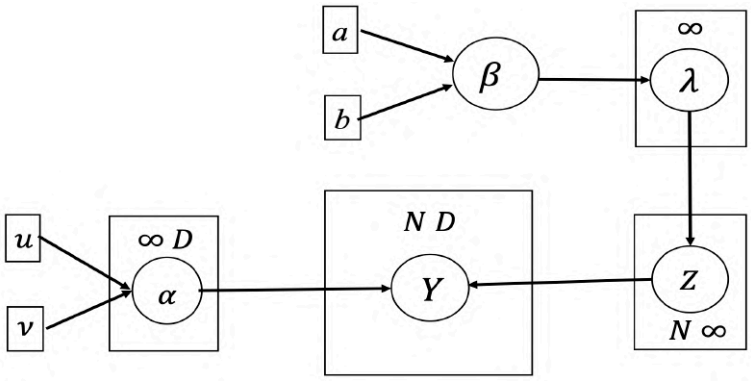


Figure 3.3: Graphical representation of the infinite multivariate Beta mixture model. Symbols in circles and squares denote random variables and model parameters, respectively. The conditional dependencies of the variables are represented by the arcs. The number mentioned in the plates indicates the number of repetition of the contained random variables.

Thus, we can obtain:

$$Q(\Theta) = \tag{3.17}$$
$$Q(\mathcal{Z})Q(\vec{\lambda})Q(\vec{\beta})Q(\vec{\alpha}) = \left[\prod_{i=1}^{N}\prod_{j=1}^{M} Q(Z_{ij})\right]\left[\prod_{j=1}^{M} Q(\lambda_j)Q(\beta_j)\right]\left[\prod_{j=1}^{M}\prod_{d=1}^{D} Q(\alpha_{jd})\right]$$

The optimal solutions for each variational posterior factor are:

$$Q(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{3.18}$$

$$Q(\vec{\lambda}) = \prod_{j=1}^{M} Beta(\lambda_j \mid c_j^*, d_j^*) \tag{3.19}$$

$$Q(\vec{\beta}) = \prod_{j=1}^{M} \mathcal{G}(\beta_j \mid a_j^*, b_j^*) \tag{3.20}$$

$$Q(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{d=1}^{D} \mathcal{G}(\alpha_{jd} \mid u^*_{jd}, \nu^*_{jd}) \tag{3.21}$$

where,

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M} \tilde{r}_{ij}} \tag{3.22}$$

$$\tilde{r}_{ij} = exp\Bigg\{ \tilde{R}_j + \sum_{d=1}^{D} \big(\overline{\alpha}_{jd} - 1\big) \ln y_{id} - \sum_{d=1}^{D} \big(\overline{\alpha}_{jd} + 1\big) \ln(1 - y_{id}) - \\ \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{d=1}^{D} \frac{y_{id}}{(1 - y_{id})}\Big] + \langle \ln \lambda_j \rangle + \sum_{s=1}^{j-1} \langle \ln (1 - \lambda_s) \rangle \Bigg\} \tag{3.23}$$

where $\tilde{R}_j$ is the approximated lower bound of R_j [135] and $R_j = \left\langle \ln \frac{\Gamma\left(\sum_{d=1}^{D} \alpha_{jd}\right)}{\prod_{d=1}^{D} \Gamma\left(\alpha_{jd}\right)} \right\rangle$ and $\tilde{R}_j$ is [135]:

$$\tilde{R}_j = \ln \frac{\Gamma\big(\sum_{d=1}^{D} \overline{\alpha}_{jd}\big)}{\prod_{d=1}^{D} \Gamma\big(\overline{\alpha}_{jd}\big)} + \tag{3.24}$$
$$\sum_{d=1}^{D} \overline{\alpha}_{jd} \left[\psi\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) - \psi\big(\overline{\alpha}_{jd}\big) \right] \times \Big[\big\langle \ln \alpha_{jd} \big\rangle - \ln \overline{\alpha}_{jd} \Big]$$
$$+ \frac{1}{2} \sum_{d=1}^{D} \overline{\alpha}^2_{jd} \left[\psi'\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) - \psi'\big(\overline{\alpha}_{jd}\big) \right] \times \Big\langle \big(\ln \alpha_{jd} - \ln \overline{\alpha}_{jd} \big)^2 \Big\rangle$$
$$+ \frac{1}{2} \sum_{a=1}^{D} \sum_{b=1, a \neq b}^{D} \overline{\alpha}_{ja}\, \overline{\alpha}_{jb} \left[\psi'\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) \times \right.$$
$$\left. \Big(\big\langle \ln \overline{\alpha}_{ja} \big\rangle - \ln \overline{\alpha}_{ja} \Big) \times \Big(\big\langle \ln \overline{\alpha}_{jb} \big\rangle - \ln \overline{\alpha}_{jb} \Big) \right]$$

The following formulas indicate the expected values:

$$u_{jd}^{*} = u_{jd} + \sum_{i=1}^{N} \langle Z_{ij} \rangle \overline{\alpha}_{jd} \Bigg[\psi\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) - \psi\big(\overline{\alpha}_{jd}\big) + \sum_{k \neq d}^{D} \psi'\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) \times \overline{\alpha}_{jd} \Big(\big\langle \ln \alpha_{jd} \big\rangle - \ln \overline{\alpha}_{jd} \Big) \Bigg] \tag{3.25}$$

$$\nu_{jd}^{*} = \nu_{jd} - \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln y_{id} - \ln(1 - y_{id}) - \ln \Big[1 + \sum_{d=1}^{D} \frac{y_{id}}{(1 - y_{id})} \Big] \Bigg] \tag{3.26}$$

$$c_{j}^{*} = 1 + \sum_{i=1}^{N} \langle Z_{ij} \rangle \,, \quad d_{j}^{*} = \langle \beta_{j} \rangle + \sum_{i=1}^{N} \sum_{s=j+1}^{M} \langle Z_{is} \rangle \tag{3.27}$$

$$a_{j}^{*} = a_{j} + 1, \quad b_{j}^{*} = b_{j} - \langle \ln (1 - \lambda_{j}) \rangle \tag{3.28}$$

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

$$\big\langle Z_{ij} \big\rangle = r_{ij}, \quad \bar{\alpha}_{jd} = \langle \alpha_{jd} \rangle = \frac{u_{jd}}{v_{jd}}, \quad \langle \beta_{j} \rangle = \frac{a_{j}^{*}}{b_{j}^{*}} \tag{3.29}$$

$$\langle \ln \lambda_{j} \rangle = \Psi\left(c_{j}^{*}\right) - \Psi\left(c_{j}^{*} + d_{j}^{*}\right), \quad \langle \ln (1 - \lambda_{j}) \rangle = \Psi\left(d_{j}^{*}\right) - \Psi\left(c_{j}^{*} + d_{j}^{*}\right) \tag{3.30}$$

$$\langle \ln \alpha_{jd} \rangle = \Psi\left(u_{jd}^{*}\right) - \ln v_{jd}^{*}, \tag{3.31}$$

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

$$\langle \lambda_{j} \rangle = \frac{c_{j}^{*}}{c_{j}^{*} + d_{j}^{*}} \tag{3.32}$$

The algorithm of variational learning of infinite multivariate Beta mixture models (MBMM) is as follows:

Algorithm 3 Variational learning of infinite MBMM.

1. Choose the initial truncation level M.
2. Initialize the values for hyperparameters u, v, a, b, c, d
3. Initialize the values of r_{ij} by K-means algorithm.
4. repeat
5. The variational E-step: Estimate the expected values by equations (3.29) to (3.31).
6. The variational M-step: Update the variational solutions using equations (3.18) to (3.21).
7. until Convergence criterion is reached.
8. Compute $\langle \lambda_j \rangle$ and substitute in equation (3.4) to obtain the estimated values of the mixing coefficients.
9. Detect the optimal number of M by eliminating the components with small mixing coefficients close to 0.

3.4 Online variational learning of DP mixtures of multivariate Beta distributions

In this section, we extend conventional variational inference to online settings following the framework in [136]. This method is a capable alternative to handle real-world situations where data points arrive in an online manner. Here, we assume that a specific amount of data is observed by t. Thus, the lower bound for this amount of observed data is maximized and calculated

as follows [137], [138], [139]:

$$\mathcal{L}^{(t)}(q) = \frac{N}{t}\sum_{i=1}^{t}\int q(\Lambda)\mathrm{d}\Lambda \sum_{\vec{Z}_i} q(\vec{Z}_i)\ln[\frac{p(\vec{Y}_i,\vec{Z}_i \mid \Lambda)}{q(\vec{Z}_i)}] + \int q(\Lambda)\ln[\frac{p(\Lambda)}{q(\Lambda)}]\mathrm{d}\Lambda \tag{3.33}$$

where $\Lambda = \{\vec{\lambda}, \vec{\alpha}\}$ and r indicates the amount of data which are currently observed. Given that $\{\vec{Y}_1, \ldots, \vec{Y}_{(t-1)}\}$ have been already observed and a new data point $\vec{Y}_t$ arrives, the current lower bound $\mathcal{L}^{(t)}(Q)$ corresponding to $Q(\vec{Z}_t)$ is maximized and updated, while fixing the other variational factors to their value at $t-1$. Thus, $Q(\vec{Z}_t)$ is defined by:

$$Q(\vec{Z}_t) = \prod_{j=1}^{M} \rho_{jt}^{Z_{jt}} \tag{3.34}$$

$$\rho_{jt} = \frac{\tilde{\rho}_{jt}}{\sum_{j=1}^{M}\tilde{\rho}_{jt}} \tag{3.35}$$

$$\begin{aligned}\tilde{\rho}_{jt} = \exp\Bigg[\tilde{R}_j + \sum_{d=1}^{D}\big(\overline{\alpha}_{jd}-1\big)\ln y_{td} - \sum_{d=1}^{D}\big(\overline{\alpha}_{jd}+1\big)\ln(1-y_{td}) \\ - \mid \overline{\alpha}_j \mid \ln\Big[1+\sum_{d=1}^{D}\frac{y_{td}}{(1-y_{td})}\Big] + \langle \ln \lambda_j^{(t-1)}\rangle + \sum_{s=1}^{j-1}\langle \ln(1-\lambda_s^{(t-1)})\rangle\Bigg]\end{aligned} \tag{3.36}$$

$\tilde{R}_j$ is defined by equation (3.24). Then, we maximize $\mathcal{L}^{(r)}(Q)$ with respect to $Q^{(t)}(\vec{\alpha})$ and $Q^{(t)}(\vec{\lambda})$, while $Q(\vec{Z}_t)$ is fixed. Thus, $Q^{(t)}(\vec{\alpha})$ and $Q^{(t)}(\vec{\lambda})$ are updated as follows:

$$Q^{(t)}(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{d=1}^{D}\mathcal{G}\big(\alpha_{jd}^{(t)} \mid u_{jd}^{*(t)}, \nu_{jd}^{*(t)}\big), Q^{(t)}(\vec{\lambda}) = \prod_{j=1}^{M} Beta\big(\lambda_j^{(t)} \mid c_j^{*(t)}, d_j^{*(t)}\big) \tag{3.37}$$

where the hyperparameters are expressed by:

$$u_{jd}^{*(t)} = u_{jd}^{*(t-1)} + \rho_t \Delta u_{jd}^{*(t)}, \quad \nu_{jd}^{*(t)} = \nu_{jd}^{*(t-1)} + \rho_t \Delta \nu_{jd}^{*(t)} \tag{3.38}$$

$$c_{jd}^{*(t)} = c_{jd}^{*(t-1)} + \rho_t \Delta c_{jd}^{*(t)}, \quad d_{jd}^{*(t)} = d_{jd}^{*(t-1)} + \rho_t \Delta d_{jd}^{*(t)} \tag{3.39}$$

where ρ_t is the learning rate defined by $\rho_t = (\tau + t)^{-\epsilon}$ [140] with two constraints, $\epsilon \in (0.5, 1]$ and $\tau \geq 0$. The main role of ρ_t is applied to decrease the effects of the earlier inaccurate inference and assists to converge faster. $\Delta u_{jd}^{*(t)}$, $\Delta \nu_{jd}^{*(t)}$, $\Delta c_{jd}^{*(t)}$ and $\Delta d_{jd}^{*(t)}$ are the natural gradient of the hyperparameters:

$$\begin{aligned}\Delta u_{jd}^{*(t)} = u_{jd} + Nr_{tj} \sum_{i=1}^{N} \langle Z_{ij} \rangle \overline{\alpha}_{jd} \Bigg[\psi\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) - \psi\big(\overline{\alpha}_{jd}\big) \\ + \sum_{k \neq d}^{D} \psi'\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) \times \overline{\alpha}_{jd} \Big(\big\langle \ln \alpha_{jd} \big\rangle - \ln \overline{\alpha}_{jd} \Big) \Bigg] - u_{jd}^{*(t-1)}\end{aligned} \tag{3.40}$$

$$\begin{aligned}\Delta \nu_{jd}^{*(t)} = \nu_{jd} + Nr_{tj} \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln y_{id} - \ln(1 - y_{id}) - \\ \ln \Big[1 + \sum_{d=1}^{D} \frac{y_{id}}{(1 - y_{id})} \Big] \Bigg] - \nu_{jd}^{*(t-1)}\end{aligned} \tag{3.41}$$

$$\Delta c_{jd}^{*(t)} = 1 + Nr_{tj} - c_{jd}^{*(t-1)}, \quad \Delta d_{jd}^{*(t)} = \psi_j + N \sum_{s=j+1}^{M} - d_j^{*(t-1)} \tag{3.42}$$

$$\begin{aligned}u_{jd}^{*} = u_{jd} + \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \overline{\alpha}_{jd} \Bigg[\psi\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) - \psi\big(\overline{\alpha}_{jd}\big) \\ + \sum_{k \neq d}^{D} \psi'\Bigg(\sum_{d=1}^{D} \overline{\alpha}_{jd} \Bigg) \times \overline{\alpha}_{jd} \Big(\big\langle \ln \alpha_{jd} \big\rangle - \ln \overline{\alpha}_{jd} \Big) \Bigg]\end{aligned} \tag{3.43}$$

$$\nu_{jd}^{*} = \nu_{jd} - \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln y_{id} - \ln(1 - y_{id}) - \ln \Big[1 + \sum_{d=1}^{D} \frac{y_{id}}{(1 - y_{id})} \Big] \Bigg] \tag{3.44}$$

$$c_j^* = 1 + \sum_{i=1}^{N} \langle Z_{ij} \rangle \,, \quad d_j^* = \langle \beta_j \rangle + \sum_{i=1}^{N} \sum_{s=j+1}^{M} \langle Z_{is} \rangle \tag{3.45}$$

The online variational learning algorithm of Dirichlet process mixture of multivariate Beta distributions is summarized in Algorithm 4.

Algorithm 4 Online variational learning algorithm.

1. Choose the initial truncation level M.
2. Initialize the values for hyperparameters.
3. for $t = 1 \rightarrow N$ do
4. The variational E-step:
5. Update the variational solution for $Q(\vec{Z}_r)$ using equation (3.34).
6. The variational M-step:
7. Compute learning rate by $\rho_r = (\tau + r)^{-\xi}$.
8. Calculate natural gradients using equation (3.40) to (3.42).
9. Update the variational solutions using equations (3.37).
10. Repeat the variational E-step and M-step until new data is observed.
11. end for

3.5 Results and discussion

In this section, we validate the performance of our proposed algorithm on three real-world medical tasks, namely, skin lesion analysis, leukemia detection and bone tissue analysis. Considering the nature of multivariate Beta distribution, we normalized all datasets in preprocessing step. Then, we applied SIFT [141] and Bag of visual words for feature extraction step of images. We applied SIFT because the SIFT feature descriptor is invariant to uniform scaling, orientation, illumination changes, and partially invariant to distortion. The 128-feature vector of each image is created by a 16x16

neighborhood around the key point and dividing it into 16 sub-blocks of 4x4 size. For each sub-block, 8 bin orientation histogram is created. So a total of 128 bin values are available. We compared our proposed models with similar algorithms and present the result in comparison tables and measure their performances in terms of accuracy, precision, recall and F1-score based on confusion matrices. The models in these tables are presented by following abbreviations: online variational learning of Dirichlet process of MB distributions (OVDPMB), batch variational learning of Dirichlet process of MB distributions (BVDPMB), online variational learning of MB mixture (OVMB) and online variational learning of Gaussian mixture models (OGMM). It should be mentioned that the computational complexity for the proposed in online and batch variational infinite mixture model is $\mathcal{O}$(MD) and $\mathcal{O}$(NMD), respectively.

3.5.1 Skin lesion analysis

Skin lesion is a serious disease with 1.04 million reported cases and ranked as the 5th typical cancer by WHO in 2018 [142]. Similar to other cases of cancer, malignant melanoma is caused by abnormal multiplication of cells and it can diffuse to other parts of the body. Therefore, early detection has a great role in increasing the survival rate and its dangerous nature can overshadowed. This cancer is primarily analyzed and diagnosed visually. However, its accurate recognition is exceedingly challenging due to similarities between skin and lesions, positive and negative cases and a wide range of physical characteristics of skin such as colors and texture. Thus, final decision and diagnosis is based on histopathological analysis of a biopsy sample. To achieve an accurate detection and precisely distinguish between melanoma and non-melanoma lesions, the pathologists should be well-trained considering the complexity of samples. To tackle this issue and detect lesions at initial stages, various Computer-Aided Diagnosis (CAD) tools [143] were applied to assist pathologists in the interpretation of biopsy images [144]. In fact due to the recent considerable advances of machine learning (ML) algorithms, CAD can integrate ML computer vision models with image processing for yielding higher accuracy in diagnosis. In our study, we applied our models on a publicly available dataset [145] which includes 726 benign and 173 malignant skin lesions. Some images of the dataset are illustrated in Figure 3.4.

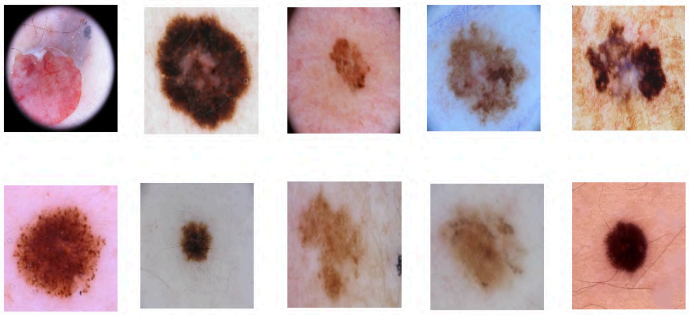


Figure 3.4: Sample of skins. The malignant and benign cases are presented in the first and second row, respectively.

The confusion matrices in Figure 3.5 and results presented in Table 3.1 illustrate the potential of our proposed model performance in this application.

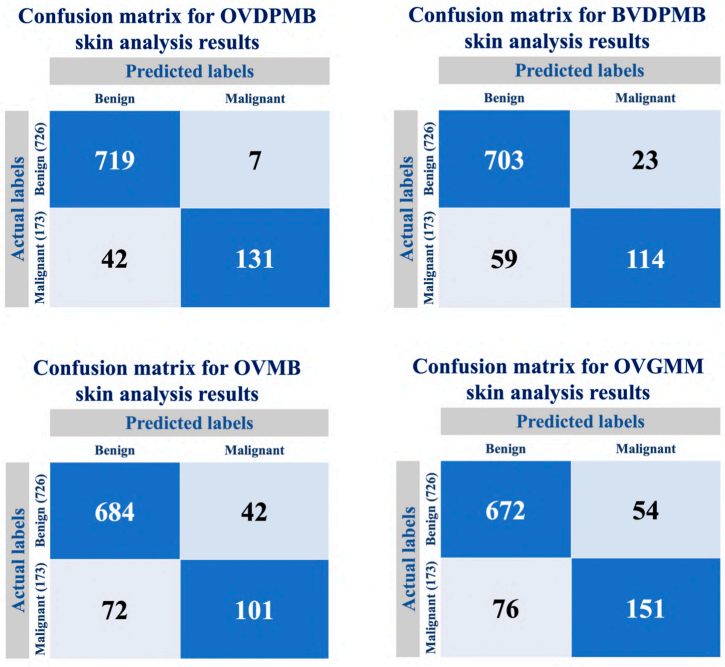


Figure 3.5: Confusion matrices for skin lesion analysis.

Table 3.1: Model performance accuracy in skin analysis.

Method	Accuracy	Precision	Recall	F1-score
OVDPMB	94.54	94.48	98.03	96.7
BVDPMB	90.87	92.25	96.83	94.48
OVMB	87.31	90.47	94.21	92.3
OVG	85.53	89.83	92.56	91.18

3.5.2 Leukemia detection

Leukemia is a fatal disease which is associated to the white blood cells (WBC). This aggressive cancer affects the spongy tissue inside the bones, called bone marrow, which is responsible to develop blood cells. This illness can results in weakening the immune system of our body. There are two principle types of leukemia, namely, acute and chronic depending on the speed of disease progression. In the former one, the WBCs can not act normally while in chronic cases they perform similar to healthy and normal blood cells. This makes it challenging to differentiate chronic cases from normal ones and there is a probability that cancerous cases get severe. There are also some subtypes such as acute Lymphocytic Leukemia, chronic Lymphocytic Leukemia, acute Myeloid Leukemia and chronic Myeloid Leukemia. Microscopic blood tests are recognized as the typical procedure for the leukemia diagnosis. Therefore, identification of leukemia existence and its specific type is a crucial task for hematologists. In fact, specification of the correct therapy for each patient and avoiding any risk depends on the accurate analysis of blood cells.
Thus, applying machine learning algorithms and intelligent tools will facilitate and accelerate analyzing blood smears or bone marrows and help to identify leukemia from the healthy samples. As we explained before, we apply a clustering method to detect leukemia as this method doesn't need annotation. Moreover, its performance is not dependent to quantity and quality of data [146]. To evaluate our approach, we used a publicly available dataset [147] which contains 600 benign and 600 malignant samples. Figure 3.6 illustrates 5 samples of each cluster.

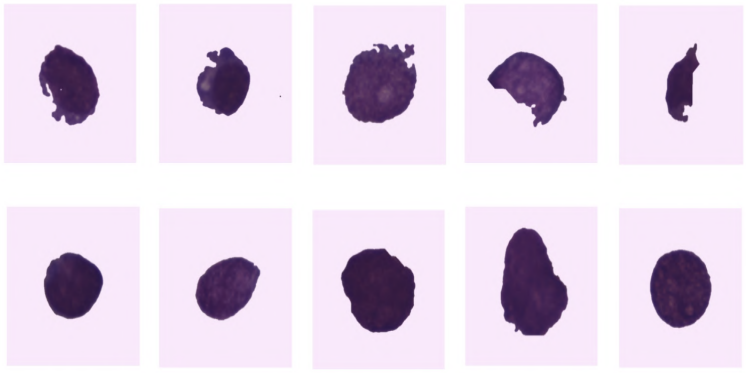


Figure 3.6: Examples of Leukemia dataset. The malignant and benign cases are presented in first and second row, respectively.

The outputs of comparing our proposed models with other alternative models are shown in Figure 3.7 and Table 3.2 which demonstrate the capability of our model.

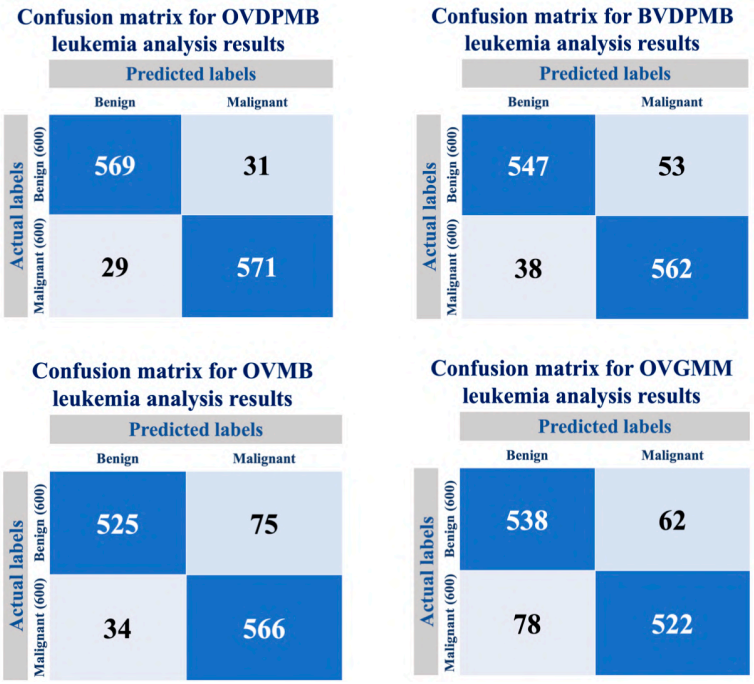


Figure 3.7: Confusion matrices for bone tissue analysis.

Table 3.2: Model performance accuracy in Leukemia analysis.

Method	Accuracy	Precision	Recall	F1-score
OVDPMB	95	95.15	94.83	94.99
BVDPMB	92.41	93.5	91.16	92.32
OVMB	90.91	93.91	87.5	90.59
OVG	88.33	87.15	89.66	88.48

3.5.3 Bone tissue analysis

Osteosarcoma is known as an aggressive and fatal tumour of the skeleton. As an uncontrollable illness, it is the most well known and common type of skeleton cancer. This disease can spread very fast all over the body according to its aggressive nature. Also, some cases could be secondary cancer and are results of metastasis and migration of cancer from another organ in the body. Grading and defining tumor type in this cancer is vital for patient as it may lead to aggressive treatment regimen. Thus, examining the biopsy samples and pathological analysis is an important stage for making final decision. In this section of our study, we applied our framework to evaluate three types of tumors, namely, benign, viable and necrotic. We use a publicly available dataset of histopathology samples [148] with 536 benign, 263 necrotic and 345 viable tumor images. Some samples are shown in Figure 3.8.

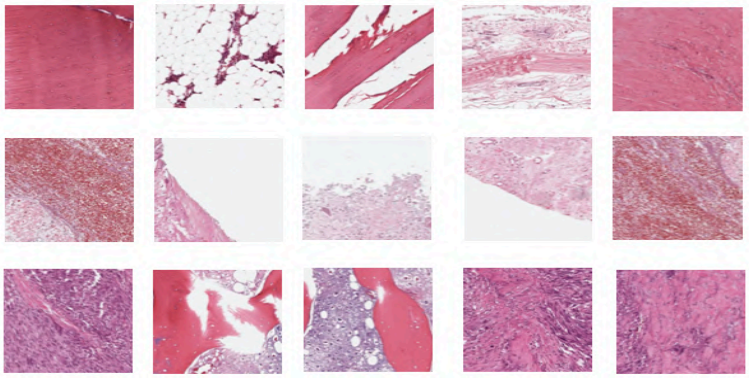


Figure 3.8: Samples of three types of bone tissues including benign, viable and necrotic tumors.

In this part, we present the comparison results in Figure 3.9 and Table 3.3. The results illustrate that our proposed framework could be considered as a capable alternative.

Table 3.3: Model performance accuracy in bone tissue analysis.

Method	Accuracy	Precision	Recall	F1-score
OVDPMB	91.09	90.99	89.88	90.43
BVDPMB	85.83	85.77	83.95	84.85
OVMB	87.5	87.1	86.55	86.82
OVG	87.93	87.12	86.83	86.97

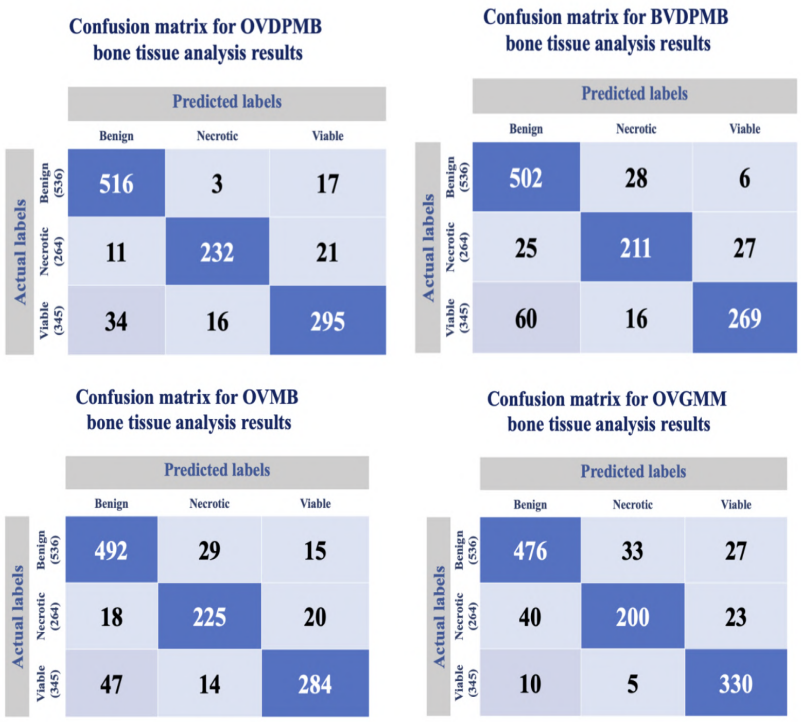


Figure 3.9: Confusion matrices for bone tissue analysis.

3.6 Conclusion

In this study, we proposed a novel clustering method, namely, Dirichlet process mixtures of multivariate Beta distributions. In the first step, we reviewed the characteristics of multivariate Beta distribution and discussed our motivation to choose it to construct our mixture model. It should be emphasized that this distribution has considerable flexibility to fit and model data. Afterward, we explained how the Dirichlet process mixture is developed. To learn our proposed model, we chose batch variational method which was explained in detail. This technique has shown robustness compared to deterministic and conventional Bayesian approach. Moreover, it can converge and estimate parameters simultaneously. To tackle the real-time cases, we extended batch learning method to an online setting. This is also a method to handle large scale datasets. We applied our novel clustering methods to medical data. The main cause to choose health-related datasets is the nature of these types of data. For instance, they are complex and heterogeneous. Another problem is highly expensive annotation which is just done by professional physicians. The other needs in healthcare cases are explainability and interpretability. All these characteristics of the medical domain, motivated us to focus on clustering methods. To evaluate and measure the performance of our novel methods, we choose three applications, namely, skin lesion analysis, leukemia detection and bone tissue analysis. To the best of our knowledge, similar clustering models and learning methods have not been widely tested on these three applications. We compared the performance of our model with other similar alternatives and presented the results in confusion matrices and comparison tables. The comparisons are based on four criteria, accuracy, precision, recall and F1-score. Considering the outcomes, the potential of our models is demonstrated. As future work, we are planning to study feature selection and integrate it to our model.

Chapter 4

Batch and online variational learning of hierarchical Dirichlet process mixtures of multivariate Beta distributions in medical applications

Thanks to the significant developments in healthcare industries, various types of medical data are generated. Analyzing such valuable resources aid healthcare experts to understand the illnesses more precisely and provide better clinical services. Machine learning as one of the capable tools could assist healthcare experts in achieving expressive interpretation and making proper decisions. As annotation of medical data is a costly and sensitive task that can be performed just by healthcare professionals, label-free methods could be significantly promising. Interpretability and evidence-based decision are other concerns in medicine. These needs were our motivators to propose a novel clustering method based on hierarchical Dirichlet process mixtures of multivariate Beta distributions. To learn it, we applied batch and online variational methods for finding the proper number of clusters as well as estimating model parameters at the same time. The effectiveness of the proposed models is evaluated on three medical real applications, namely, oropharyngeal carcinoma diagnosis, osteosarcoma analysis, and white blood cell counting.

4.1 Introduction

Spurred by astonishing progress in computer abilities in terms of processing and storage, these powerful machines are now mastering variant complex tasks that would have been deemed unapproachable a few years ago. Aligned with these developments, a wealth of data is generated which encourages scientists to extract hidden information from this valuable source. Consequently, machine learning is now applied in various domains to assist scientists in analyzing data. Similarly, with surprising advances in medical technologies, widespread health records of high-volume and large-scale are generated. It should be noted that applying machine learning (ML) in medicine is a great concern in research and healthcare industries are willing to find applicable ML solutions for the healthcare market [55]. However, there are very limited algorithms that contributed effectively to clinical applications [16]. These limitations have several roots. For instance, one of the leading reasons is that the medical inference essentially needs to be understandable, predictable and explainable. Thus, there are lots of advanced algorithms that could not be applied here. For example, the structures which produce results through black boxes may not be completely trustable [17–19]. There are other concerning issues in healthcare such as limited publicly available datasets and difficulties in the annotation of data which are barriers in applying deep learning platforms. According to their natures, such algorithms are data-hungry and need annotated datasets. Moreover, the physicians infer based on various types of data such as patient history, images, videos and signals. Thus, medical data are heterogeneous and integration is a concern in deep learning [20].

Such challenges motivated us to conduct our research on mixture models as capable clustering approaches [21, 22]. While applying these unsupervised methods, it is assumed that a set of populations constitutes the data. Through such algorithms, the model is able to divide the dataset into numerous groups. In the deployment of mixture models, we need to tackle some issues such as selecting a proper fitting distribution. Gaussian distribution has been the most ordinary one to construct Gaussian mixture models which have been widely utilized [149–155]. Nevertheless, the Gaussianity could not be assumed for all datasets. Recently, other alternatives have been applied [28,33,109,156–164]. Model's complexity determination is the second challenging aspect which has been previously tackled by some criteria such as Akaike information criterion, Bayes information criterion [111],

Minimum description length and minimum message length [112, 113] which are time-consuming. In recent researches, this challenge has been handled by non-parametric frameworks as the extension of finite mixture models. Dirichlet processes (DP) [157, 165–167] and hierarchical Dirichlet processes (HDP) [168–172] are some of these non-parametric alternatives. In these extensions, multiple populated components are produced while eliminating sparse clusters.

Our next concern in the adaptation of mixture models is parameters estimation. There are several algorithms such as deterministic and Bayesian techniques to tackle this issue. In former methods, the parameters are computed through expectation maximization [173–175]. However, such methods have some drawbacks such as dependency on initialization, over-fitting and suffering from local maxima. Bayesian techniques [176–179] have been proposed to overcome the downsides of deterministic approaches. In these algorithms, some approaches such as Markov chain Monte Carlo are employed [36, 180, 181]. Similar to the previous approach, these techniques have also some sidesteps. For instance, Markov Chain Monte Carlo prone to complex computational processing.

As a response to mitigate these weaknesses, variational inference [22, 103, 182–187] was proposed. This alternative emerging Bayesian technique is a compromise between deterministic and Bayesian frameworks. In variational solution, we approximate the posterior distribution of the model by minimizing Kullback-Leibler divergence between the true posterior and an approximated distribution as an indication of its lower bound. Securing computational tractability and convergence, variational Bayes is more manageable in comparison with purely Bayesian learning and has stronger generalization performance.

Considering all the concerns discussed above, in this work we propose first batch variational learning of HDP mixtures of MB distributions. Afterwards, we will extend it to the online setting. We can describe a summary of our four folded contributions as follows:

1. We study the behaviour of MB distribution. This relatively new distribution was presented in [24, 188] which naturally provides considerable flexibility. Its high potential in modelling data was our driver to study it. Afterwards, HDP will be constructed based on MB distribution.

2. For learning the mixture model, we apply batch variational Bayesian framework to handle two tasks simultaneously and automatically which

are estimation of model parameters and determination of model complexity.

3. As an extension to batch learning, we present online setting of variational approach. This is a capable method specifically when we need to deal with sequential data and a large number of observations. This extension is attractive, particularly in real-world applications.

4. We selected three challenging medical applications, namely, oropharyngeal carcinoma diagnosis, osteosarcoma analysis and automatic white blood cell counting. To the best of our knowledge, they haven't been widely studied with similar machine learning algorithms. To demonstrate the goodness of our proposed model performance, we will compare our models with similar alternatives. In all cases, accuracy, precision, recall and F-1 scores of each model are provided.

The remainder of this article is organized as follows: Section 4.2 is dedicated to description of our models by developing HDP mixture for MB distributions. In Section 4.3, we present the learning algorithm based on batch variational inference. Afterwards, this learning method will be extended to online setting in Section 4.4. The results of evaluation of algorithms on real-world applications are reported in Section 4.5. In Section 4.6, we present some inference remarks.

4.2 Model specification

In this section, we explain development of HDP mixture model with MB distributions.

4.2.1 Hierarchical Dirichlet process

The HDP is developed based on DP which has a Bayesian hierarchy. In this structure, the base distribution of the DP is expressed by another Dirichlet process. Here, we explain this two-level hierarchical DP model mathematically as follows:

1. At the first level, let's assume that we have a dataset $\mathcal{Y}$ including M groups. We assign a Dirichlet process G_j to each group where j is an index for each cluster.

2. Then, we consider a base distribution G_0 for the indexed set of $\{G_j\}$. This base distribution is shared among all of the groups. To form the second level of the hierarchy, another DP is assigned to G_0.

These constructions are clarified by the following equations for each $j, j \in 1, \ldots, M$:

$$\begin{aligned} &\text{First level:} \\ &G_j \sim \mathrm{DP}(\lambda, G_0), \quad G_0\text{: base distribution}, \lambda\text{: concentration parameter} \\ &\text{Second level:} \\ &G_0 \sim \mathrm{DP}(\gamma, H), \quad H\text{: base distribution}, \gamma\text{: concentration parameter} \end{aligned} \tag{4.1}$$

At both levels, we have stick-breaking construction [189–191] where we consider the total weights of clusters as a unit length which breaks to an infinite number of pieces recursively such that the size of each piece has a proportional relationship with the remainder of the stick:

$$\text{First level:} \quad \sum_{k=1}^{\infty} \pi_{jt} = 1, \quad \pi_{jt}\text{: stick-breaking weights} \tag{4.2}$$

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\varpi_{jt}}, \quad \varpi_{jt} \sim G_0, \quad \pi_{jt} = \pi'_{jt} \prod_{s=1}^{t-1} (1 - \pi'_{js}), \quad \pi'_{jt} \sim \mathrm{Beta}(1, \lambda) \tag{4.3}$$

ϖ_{jt}: a set independent random variables drawn from $G_0, \delta_{\varpi_{jt}}$: an atom at ϖ_{jt}

$$\text{Second level:} \quad \sum_{k=1}^{\infty} \psi_k = 1, \quad \psi_k\text{: stick-breaking weights} \tag{4.4}$$

$$G_0 = \sum_{k=1}^{\infty} \psi_k \delta_{\Omega_k}, \quad \Omega_k \sim H, \quad \psi_k = \psi'_k \prod_{s=1}^{k-1} (1 - \psi'_s), \quad \psi'_k \sim \mathrm{Beta}(1, \gamma) \tag{4.5}$$

Ω_k: a set independent random variables drawn from H, δ_{Ω_k}: an atom of Ω_k. Then, we define a binary latent variable as indicator for each ϖ_{jt} such that:

$$W_{jtk} = \begin{cases} W_{jtk} = 1, & \text{if } \varpi_{jt} \text{ is associated with } \Omega_k \\ W_{jtk} = 0, & \text{otherwise} \end{cases} \tag{4.6}$$

Thus, we can represent ϖ_{jt} as,

$$\varpi_{jt} = \Omega_k^{W_{jtk}} \tag{4.7}$$

Consequently, $\vec{W} = (W_{jt1}, W_{jt2}, \dots)$ is distributed as follows:

$$p(\vec{W} \mid \vec{\psi}) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \psi_k^{W_j tk} \tag{4.8}$$

According to the stick-breaking construction and considering that $\vec{\psi}$ is a function of ψ', $p(W)$ is expressed by:

$$p(\vec{W} \mid \vec{\psi'}) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \Big[\psi'_k \prod_{s=1}^{k-1}(1-\psi'_s)\Big]^{W_{jtk}} \tag{4.9}$$

To describe the dataset $\mathcal{Y}$ including M groups, each observation is indexed by i and each group is shown by index j. Assuming variable θ_{ji} as a factor assigned to an observation Y_{ij}, $\vec{\theta_j} = (\theta_{ji}, \theta_{j2}, \dots)$ are distributed by the Dirichlet process G_j. We can write the likelihood function as follows where $F(\theta_{ji})$ denotes the distribution of Y_{ji} given θ_{ji}:

$$\theta_{ji} \mid G_j \sim G_j \quad Y_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \tag{4.10}$$

To construct the HDP mixture model, H as the base distribution of G_0 is considered as prior distribution for θ_{ji}. In this setting, each group is related to a mixture model. As Ω_k is shared among all G_j, the mixture components are divided among these mixture models. Considering that θ_{ji} is distributed by G_j, it takes the value ϖ_{jt} having the probability of π_{jt}. Assigning a binary indicator variable $Z_{jit} \in \{0, 1\}$ for each θ_{ji}, we can express this latent variable by:

$$Z_{jit} = \begin{cases} Z_{jit} = 1, & \text{if } \theta_{ji} \text{ is associated with component } t \\ Z_{jit} = 0, & \text{otherwise} \end{cases} \tag{4.11}$$

Thus, we have

$$\theta_{ji} = {\varpi_{jt}}^{Z_{jit}} \tag{4.12}$$

Recalling that ϖ_{jt} maps to Ω_k, we can rewrite above equation as follows:

$$\theta_{ji} = {\varpi_{jt}}^{Z_{jit}} = {\Omega_k}^{W_{jtk}Z_{jit}} \tag{4.13}$$

We can describe $\vec{Z} = (Z_{ji1}, Z_{ji2}, \dots)$ by following equation:

$$p(\vec{Z} \mid \vec{\pi}) = \prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{t=1}^{\infty} \pi_{jt}^{Z_{jit}} \tag{4.14}$$

According to the stick-breaking construction and having $\vec{\pi}$ as a function of $\vec{\pi}'$, we have:

$$p(\vec{Z} \mid \vec{\pi}') = \prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{t=1}^{\infty} \Big[\pi'_{jt}\prod_{s=1}^{t-1}(1-\pi'_{js})\Big]^{Z_{jit}} \tag{4.15}$$

4.2.2 HDP mixture of multivariate Beta distributions

In this work, we construct HDP mixture model assuming that each observation within a component is raised from a mixture of MB distributions. This distribution was introduced in [24, 188] and its considerable flexibility and goodness in modeling motivated us to choose it.
Here, we assume first to have a dataset $\mathcal{Y} = \{\vec{Y}_1, \dots, \vec{Y}_N\}$ containing N independent and identically distributed observations. $\vec{Y}_i = \big(y_{i1}, \dots, y_{iD}\big)$ as a D-dimensional vector represents a data point drawn from a MB distribution with following probability density function where $0 < y_{id} < 1$ and $\Gamma < . >$ indicates the Gamma function:

$$p(\vec{Y}_i) = \frac{\Gamma(\mid \alpha_j \mid)\prod_{d=1}^{D} y_{id}^{\alpha_{jd}-1}}{\prod_{d=0}^{D}\Gamma(\alpha_{jd})\prod_{d=1}^{D}(1-y_{id})^{(\alpha_{jd}+1)}}\Big[1+\sum_{d=1}^{D}\frac{y_{id}}{(1-y_{id})}\Big]^{-|\alpha_j|} \tag{4.16}$$

$\vec{\alpha_j} = \big(\alpha_0, \alpha_{j1}, \dots, \alpha_{jD}\big)$ is the shape parameter where $\alpha_{jd} > 0$ for $d = 0, \dots, D$ and $\mid \alpha_j \mid = \sum_{d=0}^{D}\alpha_{jd}$. Fig. 1 illustrates four examples of MB distribution and four examples of MB mixture models [30–32, 34].

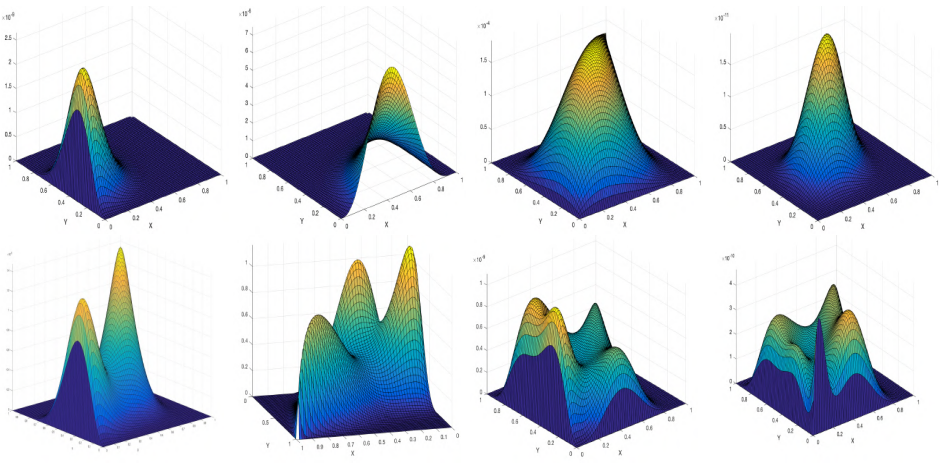


Figure 4.1: Examples of MB distribution and MB mixture models.

The likelihood function of hierarchical infinite MB mixture model taking the latent variables to account could be expressed as:

$$p(\mathcal{Y}) = \prod_{j=1}^{M} \prod_{i=1}^{N} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} p(\vec{Y}_i \mid \vec{\alpha}_k)^{Z_{jit} W_{jtk}} \tag{4.17}$$

4.3 Batch Variational Learning

Model parameter estimation is one of the main challenges when dealing with mixture models. Here, we tackle this issue with variational learning as this method has shown considerable capability. To apply this framework, we assume to have an approximation for the true posterior distribution. Here, these distributions are shown as $Q(\Theta)$ and $p(\Theta \mid \mathcal{Y})$, respectively where $\Theta = (Z, W, \psi', \pi', \alpha)$ are the set of latent and unknown random variables. To reach an optimal solution, our effort is minimizing the difference between two above-mentioned distributions using Kullback-Leibler (KL) divergence according the the following equations:

$$KL(Q \mid\mid P) = -\int Q(\Theta) \ln(\frac{p(\Theta \mid \mathcal{Y})}{Q(\Theta)}) d\Theta \tag{4.18}$$

$$KL(Q \mid\mid P) = \ln p(\mathcal{Y}) - \mathcal{L}(Q) \tag{4.19}$$

$$\mathcal{L}(Q) = \int Q(\Theta) \ln(\frac{p(\mathcal{Y}, \Theta)}{Q(\Theta)}) d\Theta \tag{4.20}$$

Recalling that $KL(Q \mid\mid P) \geq 0$, we have $KL(Q \mid\mid P) = 0$ when $Q(\Theta) = p(\Theta \mid \mathcal{Y})$. According to above-mentioned equations, $\mathcal{L}(Q) \leq \ln p(\mathcal{Y})$. Thus, $\mathcal{L}(Q)$ could be considered as a lower bound to $\ln p(\mathcal{Y})$. Due to intractability of the true posterior, we consider a computable restricted family of $Q(\Theta)$ that can be computed. With the help of mean field theory, we factorize $Q(\Theta)$ into disjoint tractable components such that:

$$Q(\Theta) = \prod_i Q_i(\Theta_i) \tag{4.21}$$

For maximizing $\mathcal{L}(Q)$, the variational methodology is applied with respect to each factor $Q_i(\Theta_i)$. For a particular Q_s, the $\{\Theta_i\}_{i \neq s}$ is fixed and $\mathcal{L}(Q)$ is maximized with respect to all forms for the distribution $Q_s(\Theta_s)$. Consequently, the optimal solution for $Q_s(\Theta_s)$ is presented by:

$$\ln Q_s(\Theta_s) = \langle \ln p(\mathcal{Y}, \Theta) \rangle_{j \neq s} + \text{const} \tag{4.22}$$

$\langle . \rangle_{j \neq s}$ shows the expectation of a particular Q_s with respect to all the distributions $Q_i(\Theta_i)$ without taking the case of $j = s$ in account, and we have:

$$\langle \ln p(\mathcal{Y}, \Theta) \rangle_{j \neq s} = \int \ln p(\mathcal{Y}, \Theta) \prod_{i \neq s} Q_i(\Theta_i) \, d\Theta_i \tag{4.23}$$

The normalized form of the solution is defined by:

$$Q_s(\Theta_s) = \frac{exp \langle \ln \ p(\mathcal{Y}, \Theta) \rangle_{j \neq s}}{\int exp \langle \ln \ p(\mathcal{Y}, \Theta) \rangle_{j \neq s} d\Theta} \tag{4.24}$$

At first step, we initialize $Q_s(\Theta_s)$ [22]. Subsequently, the factors are estimated recursively, and the value of each factor is updated in its turn using the current estimated values for the other factors. Having a convex bound guarantees the convergence of this solution [22]. Moreover, we need to define the truncation level of the stick-breaking construction. Therefore, global and group-level Dirichlet processes are set at K and T, respectively such that:

$$\psi_k{}' = 1, \quad \sum_{k=1}^{K} \psi_k = 1, \quad \psi_k = 0 \quad \text{when} \quad k > K \tag{4.25}$$

$$\pi_{jT}{}' = 1, \quad \sum_{t=1}^{T} \pi_{jt} = 1, \quad \pi_t = 0 \quad \text{when} \quad t > T \tag{4.26}$$

4.3.1 Defining prior distributions for parameters

One of the essential steps in variational learning is assigning prior distributions over model parameters. For $\vec{\alpha}$ following [192], we choose Gamma distribution denoted by $\mathcal{G}(.)$ with $\{u_{jd}\}$ and $\{v_{jd}\}$ as its positive hyperparameters:

$$p(\vec{\alpha}) = \mathcal{G}(\vec{\alpha} \mid \vec{u}, \vec{v}) = \prod_{j=1}^{\infty} \prod_{d=1}^{D} p(\alpha_{jd}) = \prod_{j=1}^{\infty} \prod_{d=1}^{D} \frac{v_{jd}^{u_{jd}}}{\Gamma(u_{jd})} \alpha_{jd}^{u_{jd}-1} e^{-v_{jd}\alpha_{jd}} \tag{4.27}$$

The prior distribution for the factors π' is given by [193]:

$$p(\pi') = \prod_{j=1}^{M} \prod_{t=1}^{\infty} Beta(1, \lambda_{jt}) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \lambda_{jt} (1 - \pi_{jt}{}')^{\lambda_{jt}-1} \tag{4.28}$$

$p(\psi')$ as the prior distribution assigned to $\vec{\psi'}$ is defined by [193]:

$$p(\psi') = \prod_{k=1}^{\infty} \text{Beta}(1, \gamma_k) = \prod_{k=1}^{\infty} \gamma_k (1 - \psi'_k)^{\gamma_k - 1} \tag{4.29}$$

4.3.2 Learning algorithm

Considering the truncated stick breaking and factorization representation, we can obtain:

$$Q(\Theta) = Q(\vec{Z})Q(\vec{W})Q(\vec{\alpha})Q(\vec{\pi}')Q(\vec{\psi}') \tag{4.30}$$

Each of the factors are represented as follows:

$$Q(\vec{Z}) = \prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{t=1}^{T}\rho_{jit}^{Z_{jit}} \tag{4.31}$$

$$Q(\vec{W}) = \prod_{j=1}^{M}\prod_{t=1}^{T}\prod_{k=1}^{K}\vartheta_{jtk}^{W_{jtk}} \tag{4.32}$$

$$Q(\vec{\pi}') = \prod_{j=1}^{M}\prod_{t=1}^{T} Beta(\pi'_{jt} \mid a_{jt}, b_{jt}) \tag{4.33}$$

$$Q(\vec{\psi}') = \prod_{k=1}^{K} Beta(\psi'_{k} \mid c_k, d_k) \tag{4.34}$$

$$Q(\vec{\alpha}) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{G}(\alpha_{kd} \mid u_{kd}, \nu_{kd}) \tag{4.35}$$

Hyperparameters of variational factors are presented in subsection 1 of Appendix B. The algorithm of variational learning of HDP mixtures of MB distributions is presented as follows where the convergence of this algorithm is guaranteed and monitored through inspection of the lower bound [22]:

Algorithm 5 Batch variational learning of HDP mixtures of MB distributions.

1. Choose the initial truncation level K and T.
2. Initialize the values of $\lambda_{jt}, \gamma_k, u_{kd}, v_{kd}$.
3. Initialize the values of ρ_{jit} by K-means algorithm.
4. repeat
5. Estimate the expected values by equations in Appendix B.
6. Update the variational solutions using equations (4.31) to (4.35).
7. until convergence criterion is reached.

4.4 Online variational learning of HDP mixtures of MB distributions

Here, we explain how to extend batch variational inference to online setting which will be used in learning of MB mixture model. Our main motivation of selecting online model is that in real life the observations arrive with an online style. To construct this model, we assume to have a particular number of data by t, and the other data points are coming. The current lower bound for the quantity of observed data by t is maximized and found by [182, 191, 194] where $\Lambda = \{\vec{W}, \vec{\psi}', \vec{\pi}', \vec{\alpha}\}$ and r indicates the amount of data which are currently observed [194]:

$$\mathcal{L}^{(r)}(q) = \frac{N}{r}\sum_{i=1}^{r}\int q(\Lambda)\mathrm{d}\Lambda \sum_{\vec{Z}_i} q(\vec{Z}_i)\ln[\frac{p(\vec{Y}_i, \vec{Z}_i \mid \Lambda)}{q(\vec{Z}_i)}] + \int q(\Lambda)\ln[\frac{p(\Lambda)}{q(\Lambda)}]\mathrm{d}\Lambda \tag{4.36}$$

The truncation procedure and factorization of $Q(\Theta)$ are similar to batch case as explained before. Assuming that $\{\vec{Y}_1, \ldots, \vec{Y}_{(r-1)}\}$ are some data points which have already been observed and $\vec{Y}_r$ as a new arrived data point, we maximize and update $\mathcal{L}^{(r)}(Q)$ as the current lower bound corresponding to

$Q(\vec{Z}_r)$ by fixing the other variational factors to their value at $r-1$. $Q(\vec{Z}_r)$ as the variational solution is expressed by:

$$Q(\vec{Z}_r) = \prod_{j=1}^{M} \prod_{t=1}^{T} \rho_{jtr}^{Z_{jtr}} \tag{4.37}$$

The calculations related to hyperparameters are presented in subsection 2 of Appendix B. Then, $\mathcal{L}^{(r)}(Q)$ is maximized with respect to $Q^{(r)}(\vec{W})$, while $Q(\vec{Z}_r)$ is fixed. Thus, we can update $Q^{(r)}(\vec{W})$ by:

$$Q^{(r)}(\vec{W}) = \prod_{j=1}^{M} \prod_{t=1}^{T} \prod_{k=1}^{K} (\vartheta_{jtk}^{(r)})^{W_{jtk}^{(r)}} \tag{4.38}$$

$$\vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r-1)} + \xi_r \Delta\vartheta_{jtk}^{(r)} \tag{4.39}$$

ξ_r as the learning rate is used to reduce the effects of the earlier inaccurate inference and helps to faster convergence. In our work, we apply a learning rate function following [140], where $\xi_r = (\tau + r)^{-\xi}$ with two constraints, $\xi \in (0.5, 1]$ and $\tau \geq 0$. $\Delta\vartheta_{jtk}^{(r)}$ as the natural gradient of $\vartheta_{jtk}^{(r)}$ is found by equation (4.40) and more details are presented in subsection 3 of Appendix B:

$$\Delta\vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r)} - \vartheta_{jtk}^{(r-1)} = \frac{\exp(\widetilde{\vartheta}_{jtk}^{(r)})}{\sum_{f=1}^{K} \exp(\widetilde{\vartheta}_{jtf}^{(r)})} - \vartheta_{jtk}^{(r-1)} \tag{4.40}$$

Then, by maximizing the current lower bound with respect to $Q^{(r)}(\vec{\pi}')$, $Q^{(r)}(\vec{\psi}')$, $Q^{(r)}(\vec{\alpha})$, we have:

$$Q^{(r)}(\vec{\pi}') = \prod_{j=1}^{M} \prod_{t=1}^{T} Beta(\pi'_{jt} \mid a_{jt}^{(r)}, b_{jt}^{(r)}) \tag{4.41}$$

$$Q^{(r)}(\vec{\psi}') = \prod_{k=1}^{K} Beta(\psi_k^{(r)} \mid c_k^{(r)}, d_k^{(r)}) \tag{4.42}$$

$$Q^{(r)}(\vec{\alpha}) = \prod_{k=1}^{K} \mathcal{G}(\alpha_k^{(r)} \mid u_k^{*(r)}, v_k^{*(r)}) \tag{4.43}$$

Hyperparameters are explained in subsection 4 of Appendix B.

Considering that hyperparameters of $Q^{(r)}(\vec{\pi}'), Q^{(r)}(\vec{\psi}'), Q^{(r)}(\vec{\alpha})$ are independent, they can be updated simultaneously. This procedure is continued till all the factors are updated with respect to the current arrived observations. To converge, the learning rate should follow these constrains:

$$\begin{cases} \sum_{r=1}^{\infty} \xi_r = \infty \\ \sum_{r=1}^{\infty} {\xi_r}^2 < \infty \end{cases} \tag{4.44}$$

The algorithm of online setting of our model is summarized in Algorithm 6.

Algorithm 6 Online variational learning of HDP mixture of MB distributions.

1. Choose the initial truncation level K and T.
2. Initialize the values for hyperparameters.
3. for $t = 1 \rightarrow N$ do
4. Update the variational solution for $Q(\vec{Z}_r)$ using equation (4.37).
5. Compute learning rate by $\xi_r = (\tau + r)^{-\xi}$.
6. Calculate natural gradients $\Delta\vartheta_{jtk}^{(r)}$ using equation in Appendix B.
7. Update the variational solutions for $Q^{(r)}(\vec{W})$ using equations (4.38).
8. Calculate the natural gradients of remaining hyperparameters using equations in Appendix B.
9. Update $Q^{(r)}(\vec{\pi}'), Q^{(r)}(\vec{\psi}'), Q^{(r)}(\vec{\alpha})$ using equations (4.41) to (4.43).
10. Repeat the variational E-step and M-step until new data is observed.
11. end for

4.5 Experimental results

In this section, we validate the performance of our proposed algorithm on three real-world medical applications, oropharyngeal carcinoma diagnosis, osteosarcoma analysis and white blood cell counting. Our motivation to choose these applications was that similar algorithms have been less studied on them. The datasets are normalized in pre-processing step and Scale-invariant feature transform (SIFT) [195] and Bag of word are employed to extract features. The models are shown by following abbreviations: online variational learning of hierarchical Dirichlet process of MB distributions "OVHDPMB", batch variational learning of hierarchical Dirichlet process of MB distributions "BVHDPMB", online variational learning of MB mixture "OVMB", online variational learning of Gaussian mixture models "OVGMM", Gaussian mixture models "GMM" and k-means. We used python for programming. We initialize the hypermarameters randomly. Based on our experience, we chose truncation level of K and T equal to 100 and 50, respectively and the values of u and v between 5 to 10. We set λ_{jt} and γ_k equal to 0.1. To assess the model performance, we used four following criteria:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{\text{Total number of observations}} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F1 - score &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned}
\tag{4.45}
$$

TP, TN, FP and FN represent the total number of true positives, true negatives, false positives, and false negatives, respectively.

4.5.1 Oral cancer diagnosis

The mouth so called oral cavity refers to the hollow section of the mouth which includes numerous parts including the lips, other sections of oropharynx such as the upper and lower gums, two thirds of the frontal part of the tongue, the area under the tongue, hard palate (the roof of mouth), the tonsils, the lining inside the lips as well as the cheeks, the walls at the back and both sides of the throat. These various sections cause this relatively small

area to have several types of tissue such as muscle, bone, nerves, rich supply blood vessels or lining called mucosa. Generally, oral diseases is one of the main health issues posing burden to numerous countries. In some cases such as cancer, it can even lead to death. Like other cases of cancer, the cells in oral cavity begin to grow out of control [196]. They could just affect any single one area or involve several parts by spreading to neighbouring tissues or even other parts of the body. In the majority of cases, oropharyngeal cancer starts in the oral cavity. This type of cancer is called oral squamous cell carcinoma (OSCC) [197, 198]. However, other cancer types like benign tumours can be also formed. OSCC is considered as the most typical malignant epithelial neoplasm [198]. Considering the statistics released by World Health Organization 4 out of 100,000 people are globally affected by this disease [199]. Despite the improvements of therapeutic approaches over the last couple of decades, morbidity and mortality rates of OSCC have not increased significantly [198]. Regardless of the easy clinical examination, the diagnosis happens in advanced stage. As noted before, several tissues and areas could be involved in this disease being unnoticed and asymptomatic at early stages. Reviewing the functions of oropharyngeal cavity, the late diagnosed or left untreated cases could be devastating. The consumption of Tobacco, alcohol and areca nut (betel quid), genetic inheritance, exposure to chemical carcinogens are key risk factors among the leading causes of OSCC [200]. However, as several cases of cancers with the help of following healthy life style, treating curable pre-malignant lesions and early stage diagnosis of precursor cases are normally very effective. To evaluate and diagnose, after preliminary examinations by caregiver, biopsy is often advised to define the cells type with pathological analysis. Based on pathology results, the OSCC lesion could be assigned to various categories for instance, metastasis, lymph node involvement or tumour. To prevent or reduce mortality rate, early detection has a great role specifically since the curable lesions are symptomatic on rare occasions. It should be emphasized that an accurate and complete pathology report is crucial to achieve precise diagnosis and design the best treatment plan for patient. As a valuable medical document in tumour staging, it includes several factors noted such as the size and shape of tumour, appearance of a specimen which are just observable with naked eye. Moreover, the extension of cancer to other parts of body is analyzed with the help of pathology. All these facts enhances the significance of analyzing the pathological samples as any error in this step can lead to irreversible damages in patient life. Thus, automation in sample

analysis could assists healthcare professionals to avoid or reduce errors. Due to dramatic improvements in computational power, novel image analysis algorithms such as computer-assisted diagnosis (CADx) allow to have digital histopathology analysis [201]. In this experiment, we evaluated our model performance accuracy by testing on a real publicly available dataset [202]. Our goal is differentiating normal epithelium and squamous cell carcinoma. This dataset includes 1224 images, including 290 benign and 934 malignant cases. Fig. 4.2 illustrates some samples of dataset.

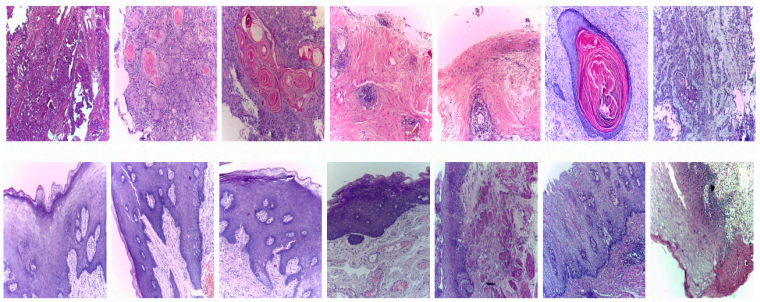


Figure 4.2: Examples of oral pathology dataset, benign and malignant cases in the first and second row, respectively.

The results in Table 4.1 show that our proposed model have good performance.

Table 4.1: Model performance accuracy in oral pathology analysis.

Method	Accuracy	Precision	Recall	F1-score
OVHDPMB	94.79	87.17	91.37	89.22
BVHDPMB	92.92	84.4	85.86	85.12
OVMB	92.84	83.22	87.24	85.18
OVGMM	90.56	77.18	85.17	80.98
GMM	89.26	75.02	81.72	78.21
K-means	88.22	72.87	79.65	76.11

4.5.2 Osteosarcoma analysis

Osteosarcoma is a primary aggressive tumour of the skeleton. It is resulted from formation a mass or lump of osteoid tissue and immature bone generated abnormally and uncontrollably by the tumour cells. This disease is the most ordinary type of bone cancers [203], starting mostly in the long bones. While benign tumours are unlikely to be fatal, they are still abnormality and can lead to future issues by compressing healthy tissue. In contrast, malignant bone tumours could grow and spread aggressively throughout the body. These cancerous tumours are sometimes secondary bone cancer, meaning that the cancer begins in another part of the body and then migrate to the bone. The root of this disease is not clear for scientist but some factors such as exposure to radiation or specific genetic changes could be associated to it. The treatment ranges from surgery to aggressive treatment regimen which depends on the grade and type of tumour. For diagnosis, beside some imaging examinations biopsy of tumour and pathological analyzing the characteristics of tissue is a critical part of staging. In this part of our experimental study, we applied our framework to asses three types of tumours, namely, benign, viable and necrotic by means of digitized histopathology samples. A publicly available dataset [148] with 536 benign, 263 necrotic and 345 viable tumour images was used in our study. Fig. 4.3 shows some samples of dataset.

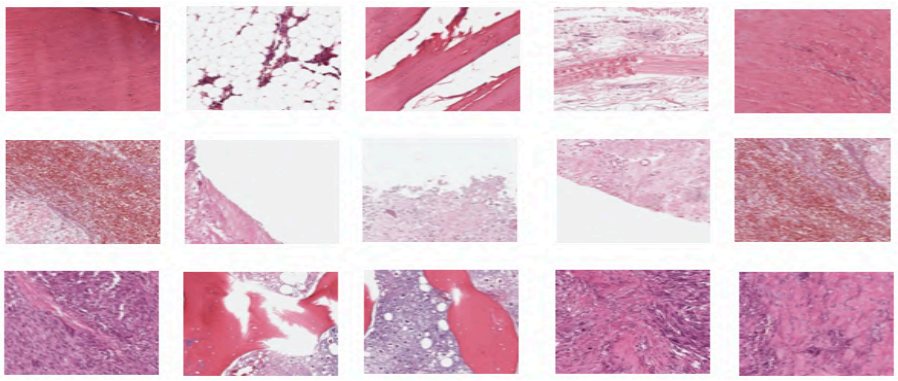


Figure 4.3: Examples of bone pathology dataset, benign, nonviable and viable cases in the first, second and third row, respectively.

The results of comparing our proposed models with other alternative models are presented in Table 4.2 which demonstrate the potential of our model as a proper alternative.

Table 4.2: Model performance accuracy in bone pathology analysis.

Method	Accuracy	Precision	Recall	F1-score
OVHDPMB	90.03	90.02	89.42	89.72
BVHDPMB	85.4	84.59	84.02	84.29
OVMB	87.5	87.1	86.55	86.82
OVGMM	87.93	87.12	86.83	86.97
GMM	86.36	85.11	85.63	85.37
K-means	85.43	84.31	84.82	84.57

4.5.3 Automatic white blood cell counting

One of the main medical diagnosis tools is complete blood cell count test which is applied to examine overall health condition by finding the amount of abnormalities in the blood smears. Traditionally this is done by manually counting the cells using some laboratory equipment. In this test, red and white cells as well as platelets are counted [204]. Being huge in number, each of these components that constitute blood has an important special role. In this work, we focus on analyzing white blood cells (WBC) which make up roughly around 1% of the whole blood volume with the approximate quantity of 4,000 to 11,000 per microliter of blood. However, they play the principle role in our immune system. Being primarily responsible to protecting the body, they defend against the infectious substances and organism which damage the biological structures. Any issue in this defending system can result in fatal diseases prevalence and even mortality. These valuable cells, called leukocytes, are composed of five components having typical share in healthy blood. These sub-classes are neutrophil with 40 to 75%, lymphocyte having 25 to 35%, monocyte with 3 to 9%, eosinophil and basophil with less than 5 and 1 percent of share, respectively. Any increase, decrease or changes in combination of WBCs could be indication of an issue in immune system or a disease. Indeed, to adequately play their role, WBCs should have sufficient amount and proper configuration of all types. Any abnormality in their total volume or proportion make a great difference to health ranging from leukemia to allergies. The traditional manual counting system could be tedious, erroneous, expensive and extremely time consuming due to the large number and variety of WBCs. There maybe also some errors in defining the

sub-types of these cells. Moreover, in term of accuracy it is vastly dependent on the expertise of the clinical laboratory professionals. Consequently, an automated procedure to count from a smear image may substantially facilitate the counting process. In essence, fast and accurate determination of WBCs distribution is critical to define the degree of abnormalities. With the distinguished improvement of machine learning approaches, image differentiation and object detection applications are achieving more robustness and accuracy. Similar to various fields of science, machine learning based analysis could be applied in this medical domain also. In this experiment, we give an account of counting four main types of WBCs through a CADx algorithm to make the analysis less prone to human error, more reliable and economical by automatic identification and counting. We applied a publicly available dataset [205] and selected a balanced configuration of four subclasses including 400 observations. Fig. 4.4 illustrates 4 examples of this dataset.

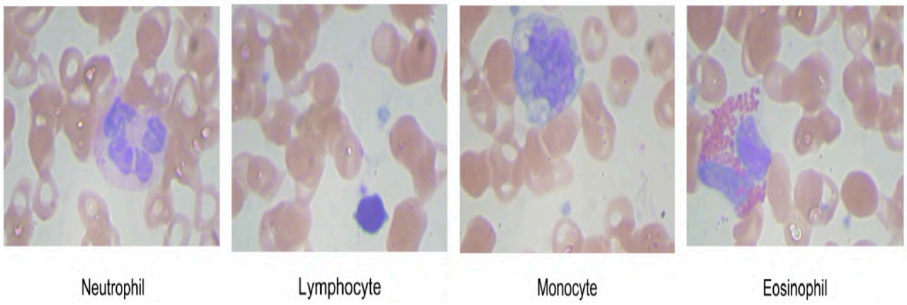


Figure 4.4: Samples of WBCs.

The outcomes of evaluating the models on WBC dataset are presented in Table 4.3 which show that our model could be considered as a capable alternative.

Table 4.3: Model performance accuracy in WBC analysis.

Method	Accuracy	Precision	Recall	F1-score
OVHDPMB	90.25	90.51	90.25	90.38
BVHDPMB	86.75	86.78	86.75	86.76
OVMB	87.25	87.71	87.25	87.48
OVGMM	87.75	87.94	86.75	87.84
GMM	86.25	86.32	86.25	86.28
K-means	86.75	86.83	86.75	86.79

4.5.4 Discussion

Considering the experimental results of three real-world applications, our proposed model has a better performance compared to other alternatives. Considering Table 4.1, OVHDPMB provides more accurate performance for oropharyngeal carcinoma diagnosis. This is supported by 94.79% of accuracy and 89.22% of F1-score. Other clustering models such as GMM and k-means have less accuracy and F1-score. BVHDPMB and OVMB have approximately similar results.

Similarly considering Table 4.2, in bone tissue analysis, OVHDPMB has 90.3% and 89.72% of accuracy and F1-score, respectively, which enhances its outperformance compared to other clustering methods. The results of OVBM and OVGMM are comparable and the lowest performance appears in K-means.

In the last part of the experiment devoted to white blood cell counting, the results in Table 4.3 show the robustness of our proposed model, OVHDPMB, as it has the highest values of accuracy, 90.25% and F1-score, 90.38%. The results of OVMB and OVGMM are similar and we got slightly close results in GMM and k-means.

Overall, these results demonstrate the potential of our proposed model as it is able to provide the highest values of accuracy and F1-score among all tested approaches. The values of the other two measurement metrics, precision, and recall also support the advantages of applying this model.

4.6 Conclusion

In this work, we proposed a novel unsupervised machine learning model, namely, HDP mixtures of MB distributions. We discussed first the characteristics of MB distribution. It should be noted that our motivation in selecting this distribution is its great flexibility in terms of fitting and modelling data. Afterwards, construction of its HDP mixtures as a flexible structure was presented. Then, the learning procedure by batch variational method was explained which is a robust technique compared to deterministic and conventional Bayesian approach. Moreover, it has the ability to converge and estimate parameters at the same time. In next step, this learning method was extended to online setting which has a significant potential in handling real-time large scale datasets and could accelerate the convergence rate in such cases.

We emphasize that applying clustering methods on medical data could be a response to complexity and heterogeneity of healthcare data and highly expensive annotation as it is just doable by health professional. Besides, physicians need to explain their inference. Thus, interpretability is another requested aspect of the model. We studied the performance of our novel method on oropharyngeal carcinoma diagnosis, osteosarcoma analysis and automatic white blood cell counting. However, we had difficulties in finding these publicly available datasets. To the best of our knowledge, similar models have not been widely tested on these three applications. At the end of each experiment, we presented comparison tables to measure the performance of models based on four criteria, accuracy, precision, recall and F1-score. The outcomes indicate the potential of our models to be considered as an alternative clustering method. We focused on computer assisted detection and analysis hoping that it assists the healthcare professionals in their error-prone tasks. This could help the healthcare system to improve quality of services provided to patients. Also, it could be beneficial in reduction or eliminate irreversible damages to patient's life. As future work, we are planning to continue our research by integrating feature selection to our model.

Chapter 5

Expectation propagation learning of finite and infinite MB mixture models

Clustering is an attractive method to handle large scale data which are explosively generated through digitization. This approach is specifically appropriate when labeling is very costly. In this paper, we constructed an unsupervised learning algorithm and focused on a finite mixture model based on multivariate Beta distribution. Our motivation is the flexibility and high potential that this distribution offers in modeling data. To learn this mixture model, we used an expectation propagation inference framework in which the parameters and the complexity of the model were evaluated concurrently in a single optimization framework. We evaluated the performance of our framework on publicly available datasets related to EEG-based sentiment analysis and human activity recognition. Our proposed model demonstrates comparable results to similar alternatives.

5.1 Introduction

As a huge amount of various types of data is generated increasingly, lots of scientists are interested to find capable approaches to analyze and manage these data and extract meaningful knowledge from these valuable resources. One of the main techniques in machine learning is clustering as an unsupervised method. This type of machine learning algorithms is specifically adequate when data labeling is tough, challenging, timely and expensive. For

instance, in healthcare, there are just medical professionals who are qualified to label healthcare data. Moreover, in such a sensitive domain unpredictable and unexplainable results generated by black boxes [17], [18], [19] are not reliable. Some algorithms such as deep learning models are data-hungry and suffer from some limitations such as issues in data integration and heterogeneity [20]. These issues motivated us to conduct our research in clustering domain. One of the most flexible and powerful clustering approaches [206], [207], [208] is finite mixture models [22], [21]. In this technique, we assume that the data includes linear combinations of limited number of a specific distributions. During last decades, Gaussian distribution has been widely used as the basic distribution to construct mixture models. However, assumption of Gaussianity can not be generalized. Recent researches have studied other alternatives such as Dirichlet [98], [156], [172], Gamma and Beta [100], [101], [102], inverted Dirichlet [106], Beta-Liouville [109]. These works have shown that other models may provide more flexibility and potential to fit non-Gaussian data. Another task while applying mixture models is defining model complexity. This issue was tackled with some selection criteria like Akaike information criterion (AIC) [110], Bayes information criterion (BIC) [111], Minimum description length (MDL) and minimum message length (MML) [112], [113]. However, these methods have some drawbacks such as being timely expensive. To learn the mixture models, various methods such as maximum likelihood [209], fully Bayesian [118] and variational [22], [210], [211], [212] methods have been proposed. Each of these techniques has its advantages and disadvantages. For instance, maximum-likelihood method via the expectation-maximization is commonly applied as it is fast in model parameter learning, but it suffers from some limitations such as convergence to a local maximum of the likelihood which affects the accuracy of model performance. Moreover, the proper number of clusters should be defined in advance. In contrast, fully Bayesian techniques are more accurate but they have high computational costs. To tackle the discussed drawbacks, some researchers have worked on an expectation propagation (EP) framework [213]. This alternative has demonstrated good performance in numerous fields. EP is a recursive approximation framework in which we try to minimize a Kullback-Leibler divergence between the true and approximated model's posterior. Another advantage of EP is that unlike maximum likelihood technique, the number of components is detected within algorithm. Thus, the model parameters and the number of components can be defined simultaneously. The major contribution of our work is as follows:

1. We construct a mixture model based on multivariate Beta distribution. The motivation behind choosing this distribution is its potential and flexibility to model the data.

2. We apply EP inference framework to estimate model parameter and model complexity simultaneously.

3. We evaluate the proposed algorithm on two applications including EEG-based sentiment analysis and human activity recognition. To evaluate the goodness of our proposed model performance, we will compare our models with other alternatives and use four metrics including accuracy, precision, recall and F1-score.

The rest of this paper is organized as follows. Section 5.2 introduces the finite multivariate Beta mixture model in details. In Section 5.3, we describe the EP inference procedure following by Section 5.4 which is devoted to adoption of EP framework for learning the multivariate Beta mixture model. Section 5.5 presents results of model evaluation on two challenging real applications. Section 5.6 closes with conclusions.

5.2 Finite multivariate Beta mixture model

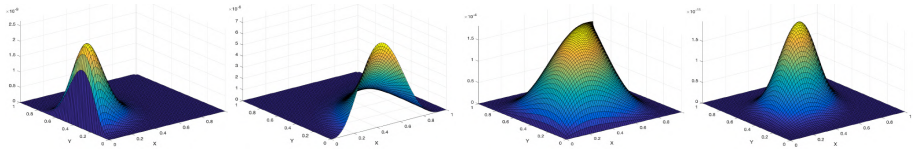
In this section, we describe finite multivariate Beta mixture models (MBMM). Olkin and Liu [24] have proposed a bivariate Beta distribution with two correlated random variables X and Y, both positive real values and less than one. The joint density function of this bivariate distribution which has three shape parameters a, b and c, is expressed as follow:

$$f(X,Y) = \frac{X^{a-1}Y^{b-1}(1-X)^{b+c-1}(1-Y)^{a+c-1}}{B(a,b,c)(1-XY)^{(a+b+c)}}, \quad B(a,b,c) = \frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(a+b+c)}.$$

For MB distribution, we define an observation by $\vec{X}_i = \big(x_{i1}, \ldots, x_{iD}\big)$ as a D-dimensional vector such that all its elements are positive and less than one. We can express the probability density function of multivariate Beta distribution [24] by Equation (5.1) where $\vec{\alpha}_j = \big(\alpha_{j0}, \ldots, \alpha_{jD}\big)$ indicates shape parameter of distribution where $\alpha_{jl} > 0$ for $l = 0, \ldots, D$ and $\mid \alpha_j \mid = \sum_{l=0}^{D} \alpha_{jl}$.

$$p(\vec{X}_i \mid \vec{\alpha}_j) = c\frac{\prod_{l=1}^{D} x_{il}^{\alpha_{jl}-1}}{\prod_{l=1}^{D}(1-x_{il})^{(\alpha_{jl}+1)}}\Big[1+\sum_{l=1}^{D}\frac{x_{il}}{(1-x_{il})}\Big]^{-|\alpha_j|}, \quad c = \frac{\Gamma(\mid \alpha_j \mid)}{\prod_{l=0}^{D}\Gamma(\alpha_{jl})}. \tag{5.1}$$

where $\Gamma(.)$ represents the Gamma function and $\mid \alpha_j \mid = \alpha_{j0} + \cdots + \alpha_{jD}$. Figure 5.1 illustrates four examples of special cases of multivariate Beta distributions (bivariate Beta distribution) and demonstrates the flexibility of this distribution.



a=1.5;b=4.5;c=7 a=4;b=1;c=4 a=2;b=2;c=2 a=4;b=5;c=4

Figure 5.1: Examples of bivariate Beta distribution with three shape parameters.

These graphs show that multivariate Beta (MB) distribution can model data with symmetric and asymmetric shapes [28, 30]. Also, in contrast with some distributions such as Dirichlet and inverted Dirichlet which are proper for proportional data and semi-bounded data, respectively, this distribution is not subjected to such constraints. To describe MB mixture model, let's assume a set of N data points which are independent and identically distributed vectors and represented by $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$. Assuming that they are generated from multivariate Beta mixture models composed of M different clusters, MB mixture model is represented by Equation (5.2). $\vec{\pi} = \big(\pi_1, \ldots, \pi_M\big)$ is the set of mixing coefficients with two constraints $\sum_{j=1}^{M}\pi_j = 1$ and $\pi_j >= 0$. $\vec{\alpha}_j$ and π_j are shape parameter and weight of component j, respectively where $j = 1, \ldots, M$.

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{M}\pi_j p\big(\vec{X}_i \mid \vec{\alpha}_j\big). \tag{5.2}$$

The likelihood function for N samples is defined as follows:

$$p\big(\mathcal{X} \mid \vec{\pi}, \vec{\alpha}\big) = \prod_{i=1}^{N}\left[\sum_{j=1}^{M}\pi_j p\big(\vec{X}_i \mid \vec{\alpha}_j\big)\right]. \tag{5.3}$$

In Figure 5.2, four examples of bivariate mixture models with 2, 3, 4 and 5 components are shown and the parameters values are presented in Table 5.1.

Table 5.1: Parameter values of bivariate mixture model plots.

Number of clusters	2	3	4	5
Values of parameter "a"	a1 = 1.5 a2 = 5	a1 = 1.41 a2 = 5.15 a3 = 5.33	a1 = 1.51 a2 = 5.11 a3 = 5.21 a4 = 1.46	a1 = 1.56 a2 = 4.8 a3 = 5.21 a4 = 1.46 a5 = 2.75
Values of parameter "b"	b1 = 4.5 b2 = 4	b1 = 4.14 b2 = 3.92 b3 = 1.42	b1 = 4.1 b2 = 4 b3 = 1.54 b4 = 8.1	b1 = 4.1 b2 = 4.1 b3 = 1.64 b4 = 8.1 b5 = 1.34
Values of parameter "c"	c1 = 7 c2 = 2.5	c1 = 7 c2 = 2.51 c3 = 8.63	c1 = 7.6 c2 = 2.52 c3 = 5.12 c4 = 3.82	c1 = 7.6 c2 = 2.58 c3 = 5.12 c4 = 3.82 c5 = 22.75

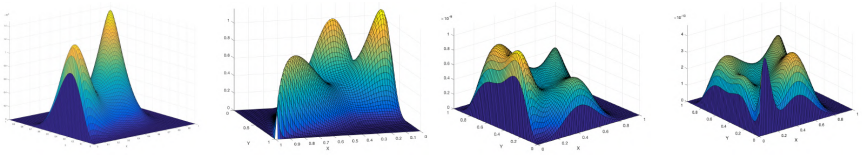


Figure 5.2: Bivariate Beta mixture models with 2, 3, 4 and 5 components.

5.3 Expectation propagation framework

In this section, we briefly present the EP approximation scheme. We assume to have a dataset including N observations which are independent and identically distributed shown as $\mathcal{X} = (\vec{X}_1, \ldots, \vec{X}_N)$. This dataset follows a model with unknown parameter. The posterior distribution is presented by $p(\Theta \mid \mathcal{X})$ where $\Theta = (\vec{\pi}, \vec{\alpha})$. In EP, the posterior distribution is approximated

with a global approximation $q(\Theta)$. The posterior distribution is defined by Equation (5.4) where $p_0(\Theta)$ and $\prod_i p_i(\vec{X}_i \mid \Theta)$ are prior distribution and contribution of each term to the likelihood, respectively.

$$p(\Theta \mid \mathcal{X}) \propto p_0(\Theta) \prod_i p_i(\vec{X}_i \mid \Theta). \tag{5.4}$$

$q(\Theta)$ as the approximating distribution should admit a factorization form similar to the true posterior. In EP, we consider one factor $f_i(\Theta) = p\left(\vec{X}_i \mid \Theta\right)$ for each data point $\vec{X}_i$ and $f_0(\Theta) = p(\Theta)$ as a factor for prior. We present the joint distribution of $\mathcal{X}$ and Θ in the form of a product of factors as follows:

$$p(\mathcal{X}, \Theta) = \prod_i f_i(\Theta). \tag{5.5}$$

Based on principle concept of EP, we approximate $p(\Theta \mid \mathcal{X})$ as the posterior distribution by a product of factors such that each factor $\widetilde{f}_i(\Theta)$ is an approximation to $f_i(\Theta)$.

$$q^*(\Theta) = \frac{\prod_i \tilde{f}_i(\Theta)}{\int \prod_i \tilde{f}_i(\Theta) d\Theta}. \tag{5.6}$$

The first step in EP learning framework is initialization of all the factors $\widetilde{f}_i(\Theta)$.

$$q^*(\Theta) = \frac{\prod_i \tilde{f}_i(\Theta)}{\int \prod_i \tilde{f}_i(\Theta) d\Theta}. \tag{5.7}$$

Afterwards, we optimize sequentially each factor in the context of the remaining factors. For a particular factor $f_j(\Theta)$, we create a cavity distribution by removing it from the current approximation to the posterior as follows:

$$q^{\setminus j}(\Theta) = \frac{q^*(\Theta)}{\widetilde{f}_j(\Theta)}. \tag{5.8}$$

Then, we obtain a new distribution which can be obtained by combining $\widetilde{f}_j(\Theta)$ with the true factor $f_j(\Theta)$:

$$\widehat{p}(\Theta) = \frac{f_j(\Theta) q^{\setminus j}(\Theta)}{\int f_j(\Theta) q^{\setminus j}(\Theta) d\Theta}. \tag{5.9}$$

We can evaluate the approximated posterior $q^*(\Theta)$ by minimizing the KL divergence $\mathrm{KL}\left(\widehat{p}(\Theta)\|q^*(\Theta)\right)$. To achieve this minimization, we match the sufficient statistics of $q^*(\Theta)$ to the corresponding moments of $\widehat{p}(\Theta)$ and update $\widetilde{f}_j(\Theta)$ as follows where $Z_j = \int f_j(\Theta)q^{\setminus j}(\Theta)d\Theta$ is a normalization constant:

$$\tilde{f}_j(\Theta) = Z_j \frac{q^*(\Theta)}{q^{\setminus j}(\Theta)}. \tag{5.10}$$

Based on EP learning framework, each factor is updated iteratively in the context of remaining factors as described in the above steps until convergence.

5.4 Expectation propagation for the multivariate Beta mixture model

In this part of our work, we apply EP framework to learn multivariate Beta mixture model. In Bayesian technique, we assign a prior distribution to each unknown parameter. For $\vec{\pi}$, we adopt a Dirichlet distribution with positive shape parameters $\vec{a} = (a_1, \ldots, a_M)$ as its conjugate prior:

$$p(\vec{\pi}) = \mathrm{Dir}(\vec{\pi} \mid \vec{a}) = \frac{\Gamma\left(\sum_{j=1}^{M} a_j\right)}{\prod_{j=1}^{M} \Gamma\left(a_j\right)} \prod_{j=1}^{M} \pi_j^{a_j - 1}. \tag{5.11}$$

For the shape parameter of multivariate Beta distribution, $\vec{\alpha}_j$, we adopt a Gaussian distribution to approximate its prior. Based on literature [214], the Gaussian provides analytically tractable calculations and can fairly capture the correlation among the elements in $\vec{\alpha}$. Thus, $\vec{\alpha}_j$ can be modeled by a D-dimensional Gaussian, with a mean vector $\vec{\mu}_j$ and a covariance matrix A_j as shown below:

$$p\left(\vec{\alpha}_j\right) = \mathcal{N}\left(\vec{\alpha}_j \mid \vec{\mu}_j, A_j\right) = \frac{|A_j|^{-1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}\left(\vec{\alpha}_j - \vec{\mu}_j\right)^T A_j^{-1}\left(\vec{\alpha}_j - \vec{\mu}_j\right)\right). \tag{5.12}$$

As we expressed in EP section, we should initialize all the approximating factors $\widetilde{f}_i(\Theta)$ at first step which is done by initializing $\{a_j, \vec{\mu}_j, A_j\}$ as the

hyperparameters. Then, the posterior approximation $q^*(\Theta)$ is initialized by setting $q^*(\Theta) \propto \prod_i \tilde{f}_i(\Theta)$. We compute the hyperparameters of $q^*(\Theta)$ as follows:

$$a_j^* = \sum_i a_{i,j} - N, \quad \vec{\mu}_j^* = \left(\sum_i A_{i,j}^{-1}\right)\left(\sum_i A_{i,j}\vec{\mu}_{i,j}\right), \quad A_j^* = \sum_i A_{i,j}. \tag{5.13}$$

To update $\tilde{f}_i(\Theta)$, we should remove it from the posterior $q^*(\Theta)$ and the corresponding hyperparameters are computed by:

$$a_j^{\setminus i} = a_j^* - a_{i,j} + 1, \quad \vec{\mu}_j^{\setminus i} = \left(A_j^{\setminus i}\right)^{-1}\left(A_j^*\vec{\mu}_j^* - A_{i,j}\vec{\mu}_{i,j}\right), \quad A_j^{\setminus i} = A_j^* - A_{i,j}. \tag{5.14}$$

The updated posterior $\widehat{p}(\Theta)$ is defined by:

$$\widehat{p}(\Theta) = \frac{1}{Z_i} f_i(\Theta) q^{\setminus i}(\Theta), \tag{5.15}$$

$$Z_i = \int f_i(\Theta) q^{\setminus i}(\Theta) d\Theta = \sum_{j=1}^{M} \frac{a_{i,j}}{\sum_j a_{i,j}} \int p\left(\vec{X}_i \mid \vec{\alpha}_j\right) N\left(\vec{\alpha}_j \mid \mu_j^{\setminus i}, A_j^{\setminus i}\right) d\vec{\alpha}_j.$$

As the integration in Equation (5.15) is not tractable, we adopt the Laplace approximation to approximate it with a Gaussian distribution [214]. So, we define a normalized distribution for this integrand which is a product of a multivariate Beta distribution and a Gaussian distribution as follows:

$$\mathcal{H}\left(\vec{\alpha}_j\right) = \frac{h\left(\vec{\alpha}_j\right)}{\int h\left(\vec{\alpha}_j\right) d\vec{\alpha}_j}, \quad h\left(\vec{\alpha}_j\right) = p\left(\vec{X}_i \mid \vec{\alpha}_j\right) \mathcal{N}\left(\vec{\alpha}_j \mid \vec{\mu}_j^{\setminus i}, A_j^{\setminus i}\right). \tag{5.16}$$

The logarithm of $h\left(\theta_{jl}\right)$ is described as follows:

$$\ln h\left(\vec{\alpha}_j\right) = \ln\Gamma(\sum_{l=0}^{D}\alpha_{jl}) - \sum_{l=0}^{D}\ln\Gamma(\alpha_{jl}) + \sum_{l=1}^{D}\alpha_{jl}\ln X_{il} \tag{5.17}$$
$$- \sum_{l=1}^{D}\ln X_{il} - \sum_{l=1}^{D}\alpha_{jl}\ln(1-X_{il}) - \sum_{l=1}^{D}\ln(1-X_{il})$$
$$- \mid \alpha_j \mid \ln(1+\sum_{l=1}^{D}\frac{X_{il}}{1-X_{il}}) - \frac{1}{2}\left(\vec{\alpha}_j - \vec{\mu}_j^{\backslash i}\right)^T A_j^{\backslash i}\left(\vec{\alpha}_j - \vec{\mu}_j^{\backslash i}\right) + \text{const.}$$

By calculating the first and second derivatives with respect to $\vec{\alpha}_j$, we have:

$$\frac{\partial \ln h(\vec{\alpha}_j)}{\partial \vec{\alpha}_j} = \begin{bmatrix} \partial \ln h\left(\vec{\alpha}_j\right)/\partial\alpha_{j0} \\ \vdots \\ \partial \ln h\left(\vec{\alpha}_j\right)/\partial\alpha_{jD} \end{bmatrix}$$
$$= \begin{bmatrix} \Psi\left(\sum_{l=0}^{D}\alpha_{jl}\right) - \Psi\left(\alpha_{j0}\right) + \ln X_{i1} - \ln(1-X_{i1}) - \ln(1+\sum_{l=1}^{D}\frac{X_{il}}{1-X_{il}}) \\ \vdots \\ \Psi\left(\sum_{l=0}^{D}\alpha_{jl}\right) - \Psi\left(\alpha_{jD}\right) + \ln X_{iD} - \ln(1-X_{iD}) - \ln(1+\sum_{l=1}^{D}\frac{X_{iD}}{1-X_{iD}}) \end{bmatrix}$$
$$-A_j^{\backslash i}\left(\vec{\alpha}_j - \vec{\mu}_j^{\backslash i}\right). \tag{5.18}$$

$$\frac{\partial^2 \ln h\left(\vec{\alpha}_j\right)}{\partial \vec{\alpha}_j^2} = \begin{bmatrix} \frac{\partial^2 \ln h(\vec{\alpha}_j)}{\partial \vec{\alpha}_{j0}^2} & \cdots & \frac{\partial^2 \ln h(\vec{\alpha}_j)}{\partial\alpha_{j1}\partial\alpha_{jD}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln h(\vec{\alpha}_j)}{\alpha_{jD}\partial\alpha_{j0}} & \cdots & \frac{\partial^2 \ln h(\vec{\alpha}_j)}{\partial\alpha_{jD}^2} \end{bmatrix} = \tag{5.19}$$
$$\begin{bmatrix} \Psi'\left(\sum_{l=0}^{D}\alpha_{jl}\right) - \Psi'\left(\alpha_{j0}\right) & \cdots & \Psi'\left(\sum_{l=0}^{D}\alpha_{jl}\right) \\ \vdots & \ddots & \vdots \\ \Psi'\left(\sum_{l=0}^{D}\alpha_{jl}\right) & \cdots & \Psi'\left(\sum_{l=0}^{D}\alpha_{jl}\right) - \Psi'\left(\alpha_{jD}\right) \end{bmatrix} - A_j^{\backslash i}.$$

$\Psi(\cdot)$ and $\Psi'(\cdot)$ are the digamma and trigamma functions. In the Laplace method, we try to find a Gaussian approximation centered on the mode of

$\mathcal{H}(\vec{\alpha}_j)$. The mode α_j^* numerically is obtained by setting the first derivative of Equation (5.18) to 0. We approximate $h(\vec{\alpha}_j)$ by its mode as follows:

$$h(\vec{\alpha}_j) \simeq h\left(\vec{\alpha}_j^*\right) \exp\left(-\frac{1}{2}\left(\vec{\alpha}_j - \vec{\alpha}_j^*\right) \widehat{A}_j \left(\vec{\alpha}_j - \vec{\alpha}_j^*\right)\right). \tag{5.20}$$

where

$$\widehat{A}_j = -\left.\frac{\partial^2 \ln h(\vec{\alpha}_j)}{\partial \vec{\alpha}_j^2}\right|_{\vec{\alpha}_j = \vec{\alpha}_j^*}. \tag{5.21}$$

Thus, the integration of $h(\vec{\alpha}_j)$ can be approximated by using Equation (5.20):

$$\begin{aligned} &\int h(\vec{\alpha}_j)\, d\vec{\alpha}_j \simeq h\left(\vec{\alpha}_j^*\right) \int \exp\left(-\frac{1}{2}\left(\vec{\alpha}_j - \vec{\alpha}_j^*\right) \widehat{A}_j \left(\vec{\alpha}_j - \vec{\alpha}_j^*\right)\right) d\vec{\alpha}_j \qquad (5.22)\\ &= h\left(\vec{\alpha}_j^*\right) \frac{(2\pi)^{D/2}}{\left|\widehat{A}_j\right|^{1/2}}. \end{aligned}$$

So, we can rewrite Equation (5.15) as follows:

$$Z_i = \sum_{j=1}^{M} \frac{a_{i,j}}{\sum_j a_{i,j}} h\left(\vec{\alpha}_j^*\right) \frac{(2\pi)^{D/2}}{\left|\widehat{A}_j\right|^{1/2}}. \tag{5.23}$$

We revise the posterior distribution $q^*(\Theta)$ by matching its sufficient statistics to the corresponding moments of $\widehat{p}(\Theta)$. So, we calculate the partial derivative of $\ln Z_i$ with respect to the model hyperparameters. For $a_j^{\setminus i}$, we have:

$$\begin{aligned} \nabla_{a_j}^{\setminus i} \ln Z_i &= \frac{1}{Z_i} \int f_i(\Theta) \frac{q^{\setminus i}(\Theta)}{q^{\setminus i}\left(\pi_j^{\setminus i}\right)} \frac{\partial}{\partial a_j^{\setminus i}} q^{\setminus i}\left(\pi_j^{\setminus i}\right) d\Theta \qquad (5.24)\\ &= \int \widehat{p}(\Theta) \left[\ln \pi_j^{\setminus i} + \Psi\left(\sum_{j=1}^{M} a_j^{\setminus i}\right) - \Psi\left(a_j^{\setminus i}\right)\right] d\Theta \\ &= E_{\widehat{p}}\left[\ln \pi_j\right] + \Psi\left(\sum_{j=1}^{M} a_j^{\setminus i}\right) - \Psi\left(a_j^{\setminus i}\right). \end{aligned}$$

With the help of moment matching, we obtain:

$$E_{\widehat{p}}\left[\ln \pi_j\right] = E_{q^*}\left[\ln \pi_j\right] = \Psi\left(a_j^*\right) - \Psi\left(\sum_{j=1}^{M} a_j^*\right). \tag{5.25}$$

We compute the partial derivatives of $\ln Z_i$ with respect to the other model hyperparameters:

$$\begin{aligned}\nabla_{\vec{\mu}_j}^{\backslash i} \ln Z_i &= \frac{1}{Z_i}\int f_i(\Theta)\frac{q^{\backslash i}(\Theta)}{q^{\backslash i}\left(\vec{\alpha}_j^{\backslash i}\right)}\frac{\partial}{\partial \vec{\mu}_j^{\backslash i}} q^{\backslash i}\left(\vec{\alpha}_j^{\backslash i}\right) d\Theta \\ &= \int \widehat{p}(\Theta)\left[A_j^{\backslash i}\vec{\alpha}_j^{\backslash i} - A_j^{\backslash i}\vec{\mu}_j^{\backslash i}\right] d\Theta = A_j^{\backslash i} E_{\widehat{p}}\left[\vec{\alpha}_j\right] - A_j^{\backslash i}\vec{\mu}_j^{\backslash i}.\end{aligned} \tag{5.26}$$

$$\begin{aligned}\nabla_{A_j}^{\backslash i} \ln Z_i &= \frac{1}{Z_i}\int f_i(\Theta)\frac{q^{\backslash i}(\Theta)}{q^{\backslash i}\left(\vec{\alpha}_j^{\backslash i}\right)}\frac{\partial}{\partial A_j^{\backslash i}} q^{\backslash i}\left(\vec{\alpha}_j^{\backslash i}\right) d\Theta \\ &= \int \hat{p}(\Theta)\left\{\frac{1}{2}\left|\left(A_j^{\backslash i}\right)^{-1}\right| - \frac{1}{2}\left[\sum_{l=1}^{D}\left(\alpha_{jl}^{\backslash i}\right)^2 - 2\alpha_{jl}^{\backslash i}\mu_{jl}^{\backslash i} + \left(\mu_{jl}^{\backslash i}\right)^2\right]\right\} d\Theta \\ &= \frac{1}{2}\left\{\left|\left(A_{jl}^{\backslash i}\right)^{-1}\right| - \left[\sum_{l=1}^{D} E_{\widehat{p}}\left[\alpha_{jl}^2\right] - 2E_{\hat{p}}\left[\alpha_{jl}\right]\mu_{jl}^{\backslash i} + \left(\mu_{jl}^{\backslash i}\right)^2\right]\right\}.\end{aligned} \tag{5.27}$$

The expectations in the above equations are obtained by the moment matching technique:

$$E_{\widehat{p}}\left[\vec{\alpha}_j\right] = E_{q^*}\left[\vec{\alpha}_j\right] = \vec{\mu}_{jl}^*, \quad E_{\widehat{p}}\left[\vec{\alpha}_j^2\right] = E_{q^*}\left[\vec{\alpha}_j^2\right] = \left(\vec{\mu}_j^*\right)^2. \tag{5.28}$$

We update the hyperparameters of $q^*(\Theta)$ by substituting the above expectations into the corresponding partial derivative equations. After obtaining $q^*(\Theta)$ and $q^{\backslash i}(\Theta)$, we update the revised hyperparameters for f_i as follows:

$$a_{i,j} = a_j^* - a_j^{\backslash i} + 1, \quad \vec{\mu}_{i,j} = A_{i,j}^{-1}\left(A_j^*\vec{\mu}_j^* - A_j^{\backslash i}\vec{\mu}_j^{\backslash i}\right), \quad A_{i,j} = A_j^* - A_j^{\backslash i}. \tag{5.29}$$

This procedure is repeated until the convergence of hyperparameters of the approximating factor. The same procedure is applied sequentially for the remaining factors. The expected values of the mixing coefficients is estimated by:

$$E\left[\pi_j\right] = \frac{a_j^*}{\sum_j a_j^*}. \tag{5.30}$$

The complete learning process is summarized in Algorithm 7.

Algorithm 7 EP learning of finite MB mixtures.

1. Choose the initial number of components M.
2. Initialize the approximating factors $\widetilde{f}_i(\Theta)$ by initializing $\{\vec{a_j}, \vec{\mu_j}, A_j\}$.
3. Initialize the posterior approximation by setting $q^*(\Theta) \propto \prod_i \tilde{f}_i(\Theta)$. The hyperparameters of $q^*(\Theta)$ are calculated by Equations (5.13).
4. **repeat**
5. Choose a factor $\widetilde{f}_i(\Theta)$ to refine.
6. Remove $\tilde{f}_i(\Theta)$ from the posterior $q^*(\Theta)$ by division $q^{\setminus i}(\Theta) = q^*(\Theta)/\tilde{f}_i(\Theta)$.
7. Evaluate the new posterior by setting the sufficient statistics (moments) of $q^*(\Theta)$ to the corresponding moments of $\widehat{p}(\Theta)$.
8. Update the factor $\widetilde{f}_i(\Theta)$ by updating the corresponding hyperparameters as in Equations (5.29).
9. **until** Convergence criterion is reached.
10. Compute the estimated values of the mixing coefficients π_j as in Equation (5.30).
11. Detect the optimal number of components M by eliminating the components with small mixing coefficients close to 0. Threshold is equal to (ϵ).

5.5 Experimental results

In this section, we evaluate our proposed model and its performance on two real-world applications, namely, EEG-based sentiment analysis and human activity recognition. Considering the nature of multivariate Beta distribution, we need to normalize our dataset in pre-processing step. The performance of our model is compared by other alternatives, variational learning of MB mixture models (VR-MBMM), variational learning of Gaussian mixture models (VR-GMM), maximum likelihood estimation of MB mixture models (ML-MBMM) and maximum likelihood estimation of Gaussian mixture models (ML-GMM). The measurement criteria are accuracy, precision, recall and F1-score. Before running our clustering algorithm, we removed the original labels and find the predicted labels by our proposed model. Then, we compared original and predicted labels. In our experiments, we chose the initial values by our experiments. We set M, $\vec{\mu}_{i,j}$, $A_{i,j}$ and $a_{i,j}$ to 20, 0.5, 0.01 and 0.2, respectively. We find the initial value of parameters arbitrary. Various sets of parameters may provide similar good accuracies in different applications and we may choose different range of parameters that works better for each of them. This part is done experimentally. As it was described in Algorithm 1, to define model complexity or number of components of mixture model, we set M to a big number and the clusters with small weights close to 0, (ϵ), will be eliminated.

5.5.1 EEG-based sentiment analysis

Discovering the association between different emotional states and specific patterns of physiological responses has attracted attention of lots of researchers and become an appealing research topic. Some studies [215] that indicate analyzing the central nervous system (CNS) provides better information associated with emotions compared to peripheral physiological responses. This belief is supported by applying neuroimaging, functional Magnetic Resonance Imaging and electroencephalographic (EEG) signals [216] to discover the relationship between brain activity and various emotional states. It should be noted that analyzing human emotion is a complex task and some emotions have overlap. Based on Lovheim Model [217], each of shame, anxiety, fear, anger, disgust, surprise, joy, and interest feelings can be mapped to generalized states of positive and negative valence. Lots of researches have been conducted to analyze emotions with the help of EEG

signals [218], [219], [220]. In this section of our experiment, we used a publicly available dataset [221] that contains EEG brainwave data collected by MUSE EEG headband with a resolution of TP9, AF7, AF8, TP10 electrodes. Labeling was performed by film clips with an obvious valence and includes positive and negative emotional states and neutral resting data [222], [223]. The participants were one male and one female individual and the data was collected for 3 minutes per state. We run our model on 2,100 observations including 700 samples for each positive, negative and neutral state. The results in Fig. 5.3 and Table 5.2 indicates the superior performance of our proposed model compared to other alternatives. As we can observe, the robustness of our proposed method is indicated by 92.23% of accuracy compared to VR-MBMM and VR-GMM with 88.19% and 85.71% of accuracy, respectively. The merit of EP inference approach and applying MB distribution over Gaussian distribution is supported with comparing the outcomes of finite mixture models.

			Predicted labels by EP-MBMM		
			Positive	Negative	Neutral
Actual labels	Positive	700	657	26	17
	Negative	700	21	630	49
	Neutral	700	27	23	650

			Predicted labels by VR-MBMM		
			Positive	Negative	Neutral
Actual labels	Positive	700	633	30	37
	Negative	700	25	618	57
	Neutral	700	45	54	601

			Predicted labels by VR-GMM		
			Positive	Negative	Neutral
Actual labels	Positive	700	608	39	53
	Negative	700	37	595	68
	Neutral	700	45	58	597

			Predicted labels by ML-MBMM		
			Positive	Negative	Neutral
Actual labels	Positive	700	596	45	59
	Negative	700	54	569	77
	Neutral	700	47	69	584

			Predicted labels by ML-GMM		
			Positive	Negative	Neutral
Actual labels	Positive	700	588	64	48
	Negative	700	72	572	56
	Neutral	700	42	76	582

Figure 5.3: Confusion matrices for EEG analysis application.

Table 5.2: Model performance accuracy in EEG-based sentiment analysis.

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
EP-MBMM	92.23	92.25	92.23	92.24
VR-MBMM	88.19	88.27	88.19	88.23
VR-GMM	85.71	85.81	85.71	85.76
ML-MBMM	83.28	83.31	83.28	83.29
ML-GMM	82.95	82.95	82.96	84.25

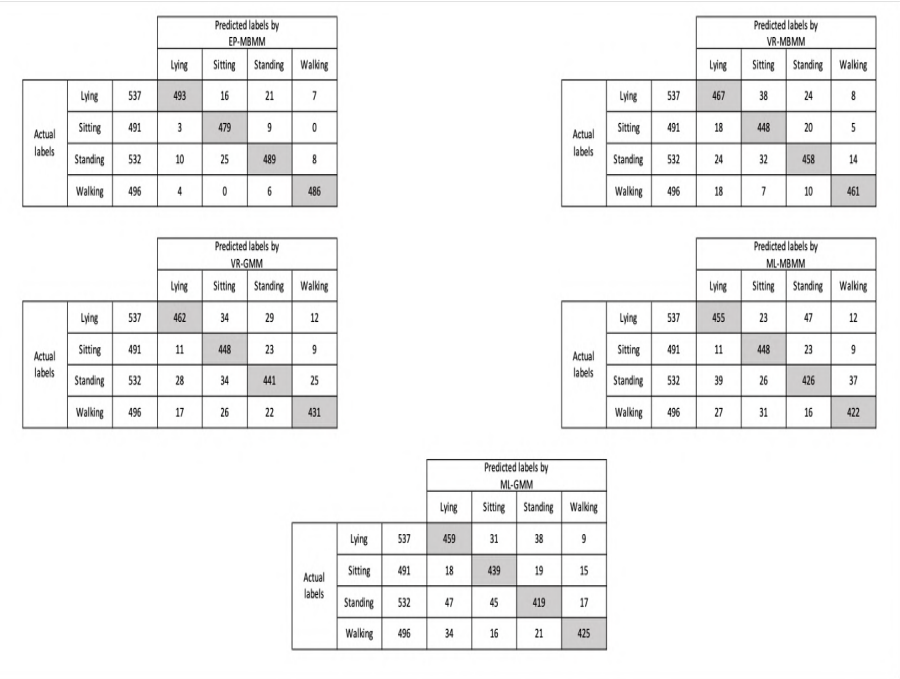
5.5.2 Physical activity recognition

In recent years, Considerable advances in information and communication technologies have resulted in broad usage of Internet of Things (IoT). However, new modern lifestyles had some consequences. Among all of the drawbacks, reduction in daily movements and having low level of physical activities led to some diseases such as morbid obesity, diabetes, cardiovascular diseases, etc. Based on WHO report, physical inactivity is ranked as the fourth leading risk factor for global mortality taking the lives of approximately 3.2 million persons. Low levels of physical activity are detrimental to the health and functioning of older people. Moreover, physical movements have positive effects on mental health and are essential for various rehabilitation plans. Thus, healthcare applications and using smart devices are becoming a part of health tracking systems. Thanks to IoT technologies, automatic and intelligent monitoring and analysis of individuals became possible. Another advantage of this branch of science is monitoring the well-being and health status of patients and being assisted-living technologies for elderly people. The data is simply collected through wearable devices and sensors. Without needing any controlled environment, the body-worn sensors or portable smart devices register activities. These cost-effective solutions retrieve valuable information from different sources, permit continuous saving of numerous signals, and during connecting to the IoT integrated system, transfer the data to a health or caregiving center. However, the nature of sensor-based data such as complexity, sensitivity to noise, unstable and temporal nature of signals result in difficulties and challenges in fulfilling the task of human activity analysis. Also, it is difficult to find a certain relationship between physical movement and generated data. Thus, several researchers focused

on developing machine learning tools to differentiate and recognize human activities. In this section of our paper, we applied our proposed model on a publicly available dataset [224], [225]. This dataset includes four types of activities, lying, sitting, standing and walking with 537, 491, 532 and 496 samples, respectively. To collect this dataset, the activities of 30 participants aged between 19 and 48 were collected while they were wearing a waist-mounted Samsung Galaxy S II smartphone. The data were recorded by two sensors, embedded accelerometer and gyroscope of smartphone and triaxial acceleration from the accelerometer (total acceleration), the estimated body acceleration, triaxial Angular velocity from the gyroscope were considered as features. To label the activities, a camera was applied. The results of comparing our proposed model with other alternative models are presented in Fig. 5.4 and Table 5.3 which demonstrates EP-MBMM provides the most accurate output with 94.69% accuracy rate in contrast with GMM which gave a less satisfying result (86.92%). Similar to previous experiment sentiment analysis, we obtain more promising performance with our proposed model. This can encourage us to conclude that EP-MBMM could be effective in this application too.

Table 5.3: Model performance accuracy in physical activity recognition.

Method	***Accuracy***	***Precision***	***Recall***	***F*1-*score***
EP-MBMM	94.69	94.74	94.78	94.76
VR-MBMM	89.34	89.35	89.43	89.39
VR-GMM	86.92	87.04	87.1	87.07
ML-MBMM	85.33	85.37	85.43	85.41
ML-GMM	84.89	85.11	84.98	85.04



Predicted labels by EP-MBMM

Actual labels			Lying	Sitting	Standing	Walking
	Lying	537	493	16	21	7
	Sitting	491	3	479	9	0
	Standing	532	10	25	489	8
	Walking	496	4	0	6	486

Predicted labels by VR-MBMM

Actual labels			Lying	Sitting	Standing	Walking
	Lying	537	467	38	24	8
	Sitting	491	18	448	20	5
	Standing	532	24	32	458	14
	Walking	496	18	7	10	461

Predicted labels by VR-GMM

Actual labels			Lying	Sitting	Standing	Walking
	Lying	537	462	34	29	12
	Sitting	491	11	448	23	9
	Standing	532	28	34	441	25
	Walking	496	17	26	22	431

Predicted labels by ML-MBMM

Actual labels			Lying	Sitting	Standing	Walking
	Lying	537	455	23	47	12
	Sitting	491	11	448	23	9
	Standing	532	39	26	426	37
	Walking	496	27	31	16	422

Predicted labels by ML-GMM

Actual labels			Lying	Sitting	Standing	Walking
	Lying	537	459	31	38	9
	Sitting	491	18	439	19	15
	Standing	532	47	45	419	17
	Walking	496	34	16	21	425

Figure 5.4: Confusion matrices for human activity recognition application.

5.6 Conclusion

In this paper, we have proposed an EP inference framework for learning finite multivariate Beta mixture models. According to the structure of this framework, model parameters and model complexity can be determined simultaneously. This procedure helps to avoid under or over-fitting. The motivation behind choosing multivariate Beta mixture models instead of commonly used mixture models such as GMM is that the assumption of Gaussianity can not be generalized. To evaluate our proposed algorithm, we conducted experiments on two real-world applications, namely, EEG-based sentiment analysis and human activity recognition. The obtained outputs are presented in comparison tables that illustrate the effectiveness and superior performance of the proposed approach. In all cases, EP-MBMM provides higher accuracy and has better results in terms of precision, recall and F1-score. Future works could be devoted to extending our work to EP learning of infinite multivariate Beta mixture model with the help of Dirichlet process to handle the difficulties in defining the proper number of mixture components.

Chapter 6

Multivariate Beta-based Hierarchical Dirichlet Process Hidden Markov Models in Medical Applications

Considering the increasing demand for analyzing sequential data in various fields of our daily lives, finding hidden patterns in a continuous flow of data is one of the interesting topics in research. Hidden Markov Models (HMMs) are one of the most powerful statistical models applied for modelling the continuous flow of new data. In this paper, we will focus on the hierarchical Dirichlet process of HMM (HDP-HMM) which has a nonparametric structure and is an advanced and elegant extension of standard parametric HMM. In the HDP-HMM process, we define transition matrices over infinite state spaces. Defining the proper number of states in HMM is one of the essential parameters which has a great impact on the inferred model. Moreover, we construct our nonparametric model based on multivariate Beta distribution. We applied variational learning approach which provides a promising strategy for inference and has been applied successfully on various domains.

6.1 Introduction

Hidden Markov Model (HMM) is a powerful approach generally applied to model Markov process systems with hidden states. This method is widely used specifically in cases where we would like to capture latent information

from observable sequential data. This method has been successfully applied in several domains of science and technology. In this chapter, we will focus on medical application of this strong modeling approach. In medicine, HMM can assist us in monitoring patient's health changes, expressing progressive alterations to patients situation or treatment process over time. For instance, it could be employed in verification of a disease development, evaluating health condition, inspecting the results and probability assessment of transitions from a healthy to a disease state. HMM could be effective in prediction and future risk estimation. There are several works devoted to HMMs such as diagnosing Schizophrenia [226], analyzing cardiac function [227–229], eye tracking [230], classification of EEG signals [231], B cell receptor sequence analysis [232], EEG-based sleep stage scoring [233], estimating dynamic functional brain connectivity [234], cancer analysis [235–238], predicting recurrence of cancers [239], genetics [240, 241], speech recognition [242–247], predicting drug response [248], cancer biomarkers detection [249], analyzing chemotherapy outcomes [250], human activity analysis [251–256] such as fall detection and senior activity analysis using motion sensors [257], HIV prediction [258], sentiment analysis [259], medical image processing [260–262], and many other applications [263–270].

However, in most of the applications, nature of sequential data is recursive. To handle this situation, some extensions to typical HMM such as hierarchical hidden Markov model [271] and hierarchical Dirichlet process hidden Markov model (HDP-HMM) [272–274] have been proposed. In particular, HDP-HMM has considerable flexibility thanks to its nonparametric structure and has been applied in various areas such as speaker diarization, abnormal activity recognition, classifying human actions, motion detection, segmentation, and classification of sequential data [275–277]. This elegant structure is a solution to one of the challenges in HMM which is defining the proper number of states. Also, it lets us learn more complex and multimodal emission distributions in the hierarchical structure of sequences in real-world applications.

Another issue while dealing with HMM is choosing a distribution for emission probabilities. In several works devoted to HMM, Gaussian Mixture Models (GMM) have been commonly used for modelling emission probability distribution [62, 278–282]. However, this assumption could not be generalized and recent researches indicate that other alternative such as Dirichlet, generalized Dirichlet, and inverted Dirichlet distribution [105, 283–285] could be considered for several types of data. Inspired by these efforts, we were motivated

to choose multivariate Beta mixture models (MBMM) which provide considerable flexibility to model symmetric, asymmetric, and skewed data [28, 30]. So, we construct our HDP-HMM model assuming that the emission probabilities follow MBMM. We call our novel HDP-HMM model "multivariate Beta-based hierarchical Dirichlet process hidden Markov models" (MB-HDP-HMM).
To learn our proposed model, a variety of approaches have been investigated. For instance, maximum likelihood approach may result in overfitting and converging towards a local maximum. Another method is fully Bayesian inference which is precise but has a long computational time. To overcome these prohibitive drawbacks, variational Bayesian approaches [182, 286, 287] have been proposed and applied to numerous machine learning algorithms. This learning method is faster than fully Bayesian one and more precise compared to the maximum likelihood approach.
Finally, we evaluate our proposed models on a medical application. The main motivation is that our model is unsupervised which makes it an adequate tool when data labelling is expensive and takes considerable time. Health-related applications are good examples because there are just medical experts who are eligible to label medical data. Moreover, having predictable and explainable results in such a sensitive domain is one of the essential needs. Therefore, decisions making and inference based on black boxes [18, 19, 288] may not be absolutely trustable. Another concerning challenge is our limitation to access a huge amount of data because of the tough confidentiality rules in healthcare. Thus, some platforms such as deep learning which provide precise results but need lots of data for learning [20] could not be easily used. Our proposed algorithm could handle datasets of various sizes, and the process is explainable in human terms.
Our contributions in this work could be summarized as follows:

1. We propose a modified version of the hierarchical Dirichlet process hidden Markov model in which emission probabilities are raised from multivariate Beta mixture models. This model, which is less costly compared to deep learning, is capable to fit different sizes of datasets and outcomes are explainable.

2. We apply variational inference to learn our proposed algorithm and secure having accurate outcomes within a proper time interval.

3. We measure the performance of our model and compare it with similar

alternatives in medical applications.

The paper is organized as follows: In section 6.2, we construct our model and describe multivariate Beta-based hidden Markov models and multivariate Beta-based hierarchical Dirichlet process of hidden Markov model. Section 6.3 is devoted to parameter estimation with variational inference. In section 6.4, we present the results of evaluating our proposed model in human activity recognition. Finally, we conclude in section 6.5.

6.2 Model Specification

To express our proposed model, we start by explaining the basic structure of HMM for a sequence of events or states. Then, we will add the assumption of having multivariate Beta mixture models as emission probabilities. We call this model, multivariate Beta-based hidden Markov model. Then, we discuss the hierarchical Dirichlet process of this modified hidden Markov model, called multivariate Beta-based hierarchical Dirichlet process hidden Markov model.

6.2.1 Multivariate Beta-based Hidden Markov Model

Further to the Markovian characteristics of HMM, in the first-order Markov model, the probability of each event t depends just on state $t-1$ which happens immediately before t. In HMM, a system with hidden states emits observable symbols at any specific point of time.
To mathematically formulate HMM, we need following parameters:
- Transition probability: indicating the probability of a change in state from t to $t+1$. Sum of all these probabilities given the current state is equal to 1.
- Initial Probability: The initial state that the system starts from it is denoted as π. These probabilities also sum up to 1.
- Emission Probability or observation likelihoods: parameters indicating the probability of a data point being generated from a specific state.
In our work, HMM is expressed by $\lambda = \{A, B, \varphi, \pi\}$ and following notations:
1. T: length of the sequence of our interest, M: number of mixture components in set $L = \{m_,..., m_M\}$, K: number of the states.
2. A state sequence $\mathcal{S} = \{S_1, \ldots, S_T\}$ drawn from $P(s_t \mid s_{t-1}, \ldots, s_1) = P(s_t \mid s_{t-1})$.
3. Sequential data $\mathcal{X} = \{X_1, \ldots, X_T\}$.

4. Transition probability from state i to i': $A = \{a_{ii'} = P(s_t = i' \mid s_{t-1} = i)\}$.
5. Emission probability of observing j from state i: $B = \{B_{ij} = P(m_t = j \mid s_t = i)\}$ for $j \in [1, M]$.
6. π_j: Initial probability to begin the sequence from the state j.
7. φ is the set of mixture parameters. In this work, we apply multivariate Beta mixture model and φ is the shape parameter, $\alpha_{ij} = (\alpha_{1ij}, \ldots, \alpha_{Dij})$, with $i \in [1, K]$ and $j \in [1, M]$.

We can denote the complete likelihood of HMMs as follows:

$$p(\vec{X} \mid A, B, \pi, \alpha) = \sum_{S} \sum_{L} \pi_{s_1} [\prod_{t=1}^{T-1} a_{s_t, s_{t+1}}] [\prod_{t=1}^{T} b_{s_t, m_t} p(x_t \mid \alpha_{s_t, m_t})] \tag{6.1}$$

Here, we explain the model for one sequence. In case of having more observations, this could be generalized by adding a summation over the whole sequence.
$p(x_t \mid \alpha_{s_t, m_t})$ is multivariate Beta distribution (MB). To describe it in detail, let's assume to have a D-dimensional observation, $\vec{X} = \big(x_1, \ldots, x_D\big)$ where all its elements are greater than zero and less than one.
The probability density function of multivariate Beta distribution [24] is expressed as follows:

$$p(\vec{X} \mid \vec{\alpha}) = \frac{\Gamma(\mid \alpha \mid) \prod_{d=1}^{D} x_d^{\alpha_d - 1}}{\prod_{d=0}^{D} \Gamma(\alpha_d) \prod_{d=1}^{D} (1 - x_d)^{(\alpha_d + 1)}} \Big[1 + \sum_{d=1}^{D} \frac{x_d}{(1 - x_d)}\Big]^{-|\alpha_j|} \tag{6.2}$$

$\vec{\alpha} = \big(\alpha_0, \ldots, \alpha_D\big)$ is shape parameter such that $\alpha_d > 0$ for $d = 0, \ldots, D$, $\mid \alpha \mid = \sum_{d=0}^{D} \alpha_d$ and $\Gamma(.)$ represents the Gamma function.
Figures 6.1 and 6.2 illustrate some examples of multivariate Beta distributions and multivariate Beta mixture models, respectively. These figures illustrate the flexibility of this distribution. So, it has the capability of capturing symmetric and asymmetric shapes of data [28, 30].

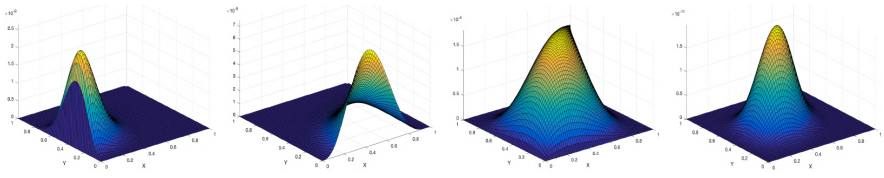


Figure 6.1: Multivariate Beta distribution with different shape parameters.

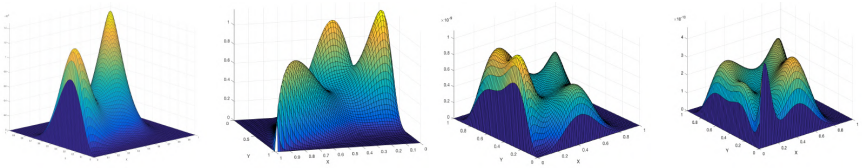


Figure 6.2: Multivariate Beta mixture models with 2, 3, 4 and 5 components.

Thus, assuming that emission probabilities are raised from MB mixture model, complete log-likelihood of $p(\vec{X} \mid A, B, \pi, \alpha)$ could be written as:

$$\log(p(\vec{X}, Z \mid \lambda)) = \log(\pi_{s_1}) + \sum_{t=1}^{T-1} \log(a_{s_t, s_{t+1}}) + \sum_{t=1}^{T} \log(b_{s_t, m_t}) +$$
$$+ \sum_{t=1}^{T} \left[\log(\Gamma(\sum_{d=0}^{D} \alpha_d)) - \log(\prod_{d=0}^{D} \Gamma(\alpha_d)) + \sum_{d=1}^{D} \Big((\alpha_d - 1) \log x_d \Big) \right.$$
$$\left. - \sum_{d=1}^{D} \Big((\alpha_d + 1) \log(1 - x_d) \Big) - (\sum_{d=0}^{D} \alpha_d) \log \left[1 + \sum_{d=1}^{D} \frac{x_d}{1 - x_d} \right] \right] \tag{6.3}$$

6.2.2 Multivariate Beta-based Hierarchical Dirichlet Process of Hidden Markov Model

To express our hierarchical HMM, we need first to describe Dirichlet process (DP) and stick breaking construction [124, 190]. The Dirichlet process [289] is an extension of the Dirichlet distribution. It has two inputs, a nonnegative precision scalar, ϵ and a base distribution G_0. DP is defined over the measurable space $(\Theta, \mathcal{B})$. For a disjoint sets of $B = \{B_1, \ldots, B_D\}$ and partition of Θ, the Dirichlet process is defined as follows where $\bigcup_i B_i = \Theta$:

$$(G(B_1), \ldots, G(B_D)) \sim \text{Dir}(\epsilon G_0(B_1), \ldots, \epsilon G_0(B_D)) \tag{6.4}$$

In terms of dimensionality, DP is infinite ($D \to \infty$). If we draw G from a DP expressed by $G \sim DP(\epsilon G_0)$, we will have:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i} \tag{6.5}$$

θ_i indicated the location drawn from G_0 and is related to a measure, p_i.

We can consider θ_i as the emission probability at state i in HMM. To move forward, we need to explain general definition of a stick-breaking process. Let's assume a probability mass function $p = (p_1, \ldots, p_{d+1})$, so we have:

$$p_i = V_i \prod_{i'=1}^{i-1} (1 - V_{i'}), \quad p_{d+1} = 1 - \sum_{i=1}^{d} p_i \quad V_i \sim \text{Beta}(v_i, \omega_i) \tag{6.6}$$

$v_i = (v_1, \ldots, v_d)$ and $\omega_1 = (\omega_d, \ldots, \omega_i)$ are non-negative, real parameters for $i = 1, \ldots, d$. The value of d could be either finite or infinite and finite case is similar to a distribution called generalized Dirichlet distribution (GDD) [290, 291]. In infinite case, we may have various ranges of priors by changing v and ω [292]. For HDP-HMM, we construct a draw from DP with following representation of a stick-breaking process:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}, \quad p_i = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad V_i \sim \text{Beta}(1, \gamma), \quad \theta_i \sim G_0 \tag{6.7}$$

$\gamma = \sum_i \beta_i$ affects a draw from DP. If $\gamma \to 0$, a measure degeneration at a random component with location drawn from G_0 happens. In contrast, if $\gamma \to \infty$, the breaks are very small and G which reach to convergence to the empirical distribution of the individual draws from G_0, and G_0 is reproduced. If we focus on a distribution from which we draw the data and show it by $p(x \mid \theta)$ with parameter θ, a DP mixture model is presented by:

$$x_i \mid \theta_i \sim p(x \mid \theta_i), \quad \theta_i \mid G \sim G, \quad G \mid \gamma G_0 \sim DP(\gamma G_0) \tag{6.8}$$

Hidden Markov models could be considered as a special case of mixture models which are dependent on the states. The supports of the mixtures are shared among them with various mixing weights. We represent state-dependent mixture model of HMM as follows where $\theta_i \equiv (b_{i1}, \ldots, b_{iM})$, distribution is MB and initial state is selected from π:

$$x_t \mid \theta_{s_t} \sim \mathcal{MB}(\theta_{s_t}), \quad \theta_{s_t} \mid s_{t-1} \sim G_{s_{t-1}}, \quad G_i = \sum_{i'=1}^{D} a_{ii'} \delta_{\theta_{i'}} \tag{6.9}$$

If we consider each transition as a DP, it will make a problem specifically if we assume that each row, i, is raised from an infinite transition probability matrix expressed as follows:

$$G_i = \sum_{i'=1}^{\infty} a_{ii'} \delta_{\theta_{ii'}}, \quad a_{ii'} = V_{ii'} \prod_{k=1}^{i'-1} (1 - V_{ik}), \quad V_{ii'} \sim \text{Beta}(1, \gamma), \quad \theta_{ii'} \sim G_0 \tag{6.10}$$

$a_{ii'}$ presents the i'^{th} component of a_i which is an infinite vector. In case of having a continues G_0, the probability of transition to a previous state is zero for each $\theta_{ii'}$ because $p(\theta_m = \theta_n) = 0$ for $m \neq n$. Thus, such approaches are not practical to construct Dirichlet process of HMM.
Hierarchical Dirichlet Process Hidden Markov Model is proposed to tackle this issue. In hierarchical Dirichlet process (HDP), the base distribution, G_0, over Θ is itself arised from a DP which relatively assure us that G_0 will be almost discrete. We formulate the process as:

$$G_m \sim DP(\beta G_0), \quad G_0 \sim DP(\gamma H) \tag{6.11}$$

In HDP as a two-level hierarchical structure, the distribution on the data points in Θ is changed from the continuous H to the discrete, but infinite G_0. If we draw for G_m multiple times, the weight on the same set of states will be substantial. This procedure and second level of DP can be expressed as follows with truncation level of K:

$$G_0 = \sum_{i=1}^{K} p_i \delta_{\theta_i}, \quad p_i = V_i \prod_{i'=1}^{i-1} (1 - V_{i'}), \quad V_i \sim \text{Beta}(1, \gamma), \quad \theta_i \sim H \tag{6.12}$$
$$(G_m(\theta_1), G_m(\theta_2), \ldots, G_m(\theta_K)) \sim \text{Dir}(\beta p_1, \beta p_2, \ldots, \beta p_K)$$

$G(\theta_i)$ indicates a probability measure at location θ_i. To summarize the procedure of two-level hierarchy, we assume to have a DP at top level through which the number of states and their observation parameters are chosen. Then, the mixing weights are considered as prior for second level where the transition probabilities are drawn. As a conjugacy between these two levels doesn't exist, there is not a truly variational solution [293]. To construct HDP-HMM, we use a prior similar to equation (6.6) which is more general

and flexible compared to the stick-breaking process for drawing from the DP, in which we draw simultaneously both of Beta$(1, \alpha)$-distributed random variables and the atoms associated with the resulting weights. As we explained before equation (6.6) could be considered as a GDD and its density function of $\boldsymbol{V} = (V_1, \ldots, V_K)$ is expresses as follows where $v = (v_1, v_2, \ldots)$ and $\omega = (\omega_1, \omega_2, \ldots)$:

$$f(\boldsymbol{V}) = \prod_{i=1}^{K} f(V_i) = \prod_{i=1}^{K} \frac{\Gamma(v_i + \omega_i)}{\Gamma(v_i)\Gamma(\omega_i)} V_i^{v_i - 1} (1 - V_i)^{\omega_i - 1} \tag{6.13}$$

By changing $\boldsymbol{V}$ to $\boldsymbol{p}$, the density of $\boldsymbol{p}$ is defined by:

$$f(\boldsymbol{p}) = \prod_{i=1}^{K} \left(\frac{\Gamma(v_i + \omega_i)}{\Gamma(v_i)\Gamma(\omega_i)} p_i^{v_i - 1} \right) p_{K+1}^{\omega_K - 1} (1 - P_1)^{\omega_1 - (v_2 + \omega_2)} \times \cdots \times \tag{6.14}$$
$$(1 - P_{K-1})^{\omega_{K-1} - (v_{K-1} + \omega_{K-1})}$$

Mean and variance for each element, p_i, is:

$$\mathbb{E}[p_i] = \frac{v_{i'} \prod_{\ell=1}^{i'-1} \omega_\ell}{\prod_{\ell=1}^{i'} (v_\ell + \omega_\ell)}, \quad \mathbb{V}[p_i] = \frac{v_{i'} (v_{i'} + 1) \prod_{\ell=1}^{i'-1} \omega_\ell (\omega_\ell + 1)}{\prod_{\ell=1}^{i'} (v_\ell + \omega_\ell)(v_\ell + \omega_\ell + 1)} \tag{6.15}$$

GDD is a special case of typical standard Dirichlet distribution. In GDD case, the construction of $\boldsymbol{p}$ from the infinite process of equation (6.14) is refered by $\boldsymbol{p} \sim GDD(\boldsymbol{v}, \boldsymbol{\omega})$. For a set of N observations which are independent identically distributed (iid) , $X_n \overset{iid}{\sim} \text{Mult}(\boldsymbol{p})$, the posterior of the respective priors presented by $\boldsymbol{v}'$ and ω' are parametrized as follows:

$$v_i' = v_i + \sum_{n=1}^{N} \mathbf{1}(X_n = i), \quad \omega_i' = \sum_{j>i} \sum_{n=1}^{N} \mathbf{1}(X_n = j) \tag{6.16}$$

$\mathbf{1}(\cdot)$ is an indicator function which will be equal to one if the argument is true and zero, otherwise. This is applied to count the number of times the random variables are equal to values of interest.

6.3 Variational Learning

To estimate model's parameters, we adopt variational inference. In this method, we introduce an approximating distribution $q(A, B, \pi, \alpha, S, L)$ for the true posterior $p(A, B, \pi, \alpha, S, L \mid \vec{X})$. Then, we try to minimize the distance between these two distributions with the help of Kullback-Leibler distance. As marginal distribution is not tractable, we try to find a tractable lower bound in it. Based on Jensen's inequality, as $KL(q \mid\mid p) \geq 0$, $KL(q \mid\mid p) = 0$ when q is equal to true posterior. $\mathcal{L}(q)$ as a lower bound to $\ln p(\vec{X})$ could be found by:

$$\ln(p(\vec{X})) = \mathcal{L}(q) - \mathrm{KL}(q(A, B, \pi, \alpha, S, L) \| p(A, B, \pi, \alpha, S, L \mid \vec{X})) \quad (6.17)$$

The true posterior distribution is practically intractable and cannot be directly applied in variational inference. Borrowing the idea from mean field theory, we consider a restricted family of distributions q and adopt a factorization approach [294, 295]. So, we have:

$$q(A, B, \pi, \alpha, S, L) = q(A)q(B)q(\pi)q(\alpha)q(S, L) \quad (6.18)$$

With the help of iterative expectation maximization (EM), we perform this approximation. Expectation step is as follows [296] such that m_i is the expected number of data points from a component in an iteration with truncation to K-dimensions:

$$\langle \ln V_i \rangle = \psi\left(1 + \langle x_i \rangle\right) - \psi\left(1 + \gamma_i + \sum_{i'=i}^{K} \langle x_{i'} \rangle\right) \quad (6.19)$$

$$\langle \ln\left(1 - V_i\right) \rangle = \psi\left(\gamma_i + \sum_{i'=i+1}^{K} \langle x_{i'} \rangle\right) - \psi\left(1 + \gamma_i + \sum_{i'=i}^{K} \langle x_{i'} \rangle\right) \quad (6.20)$$

$$\langle \ln p_1 \rangle = \langle \ln V_i \rangle \quad (6.21)$$

$$\langle \ln p_k \rangle = \langle \ln V_k \rangle + \sum_{i'=1}^{k-1} \langle \ln\left(1 - V_{i'}\right) \rangle \quad 2 \leq k < K \quad (6.22)$$

$$\langle \ln p_K \rangle = \sum_{i'=1}^{K-1} \langle \ln\left(1 - V_{i'}\right) \rangle \quad (6.23)$$

ψ represents the digamma function. Then, we optimize following quantity:

$$\ln(p^{\star}(X_t \mid \alpha_{s_t,m_t})) = \phi_{ijt}^{B} \int q(\alpha) \ln(p(X_t \mid \alpha_{s_t,m_t}))d\alpha \tag{6.24}$$

$$= \phi_{ijt}^{B} \int q(\alpha) \ln(\frac{\Gamma(\sum_{d=1}^{D} \alpha_{ijl})}{\prod_{d=1}^{D} \Gamma(\alpha_{ijl})})d\alpha + \phi_{ijt}^{B} \int q(\alpha) \Bigg[\sum_{d=1}^{D} \Big((\alpha_d - 1) \log x_{td} \Big)$$

$$- \sum_{d=1}^{D} \Big((\alpha_d + 1) \log(1 - x_{td}) \Big) - (\sum_{d=0}^{D} \alpha_d) \log \Big[1 + \sum_{d=1}^{D} \frac{x_{td}}{1 - x_{td}} \Big] \Bigg] d\alpha$$

where $\phi_{ijt}^{B} \triangleq q(s_{t-1} = i, m_t = j)$ and $*$ indicates an optimized parameter. $p(X_t \mid \alpha_{s_t,m_t})$ is MB distribution. We presented in detail the variational inference of multivariate Beta mixture models in our previous works [30,192] and similar to them, we have:

$$\bar{\alpha}_{ijl} = \frac{u_{ijl}}{v_{ijl}} \tag{6.25}$$

$$\langle \frac{\Gamma(\sum_{d=1}^{D} \alpha_{ijl})}{\prod_{d=1}^{D} \Gamma(\alpha_{ijl})} \rangle = \ln \frac{\Gamma\big(\sum_{l=1}^{D} \overline{\alpha}_{jl}\big)}{\prod_{l=1}^{D} \Gamma\big(\overline{\alpha}_{jl}\big)} + \sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg[\psi\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) - \psi\big(\overline{\alpha}_{jl}\big) \Bigg]$$

$$\times \Big[\big\langle \ln \alpha_{jl} \big\rangle - \ln \overline{\alpha}_{jl} \Big] + \frac{1}{2} \sum_{l=1}^{D} \overline{\alpha}_{jl}^{2} \Bigg[\psi'\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) - \psi'\big(\overline{\alpha}_{jl}\big) \Bigg]$$

$$\times \Big\langle \big(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl} \big)^2 \Big\rangle + \frac{1}{2} \sum_{a=1}^{D} \sum_{b=1, a \neq b}^{D} \overline{\alpha}_{ja}\, \overline{\alpha}_{jb} \Bigg[\psi'\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg)$$

$$\times \Big(\big\langle \ln \overline{\alpha}_{ja} \big\rangle - \ln \overline{\alpha}_{ja} \Big) \times \Big(\big\langle \ln \overline{\alpha}_{jb} \big\rangle - \ln \overline{\alpha}_{jb} \Big) \Bigg] \tag{6.26}$$

$$\langle \ln(\alpha_{ijd}) \rangle = \Psi(u_{ijd}) - \ln(v_{ijd}) \tag{6.27}$$

In maximization step, we update variational factors as follows:

$$q(A) = \prod_{i=1}^{K} GDD(\boldsymbol{v'}_i, \boldsymbol{\omega'}_i) \tag{6.28}$$

$$q(\alpha) = \prod_{i=1}^{K}\prod_{j=1}^{M} q(\alpha_{ij}), \quad q(\alpha_{ij}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_{ijl} \mid u^{\star}_{ijl}, v^{\star}_{ijl}) \tag{6.29}$$

$$q(\pi) = \mathcal{D}(\boldsymbol{v'}_{\pi}, \boldsymbol{\omega'}_{\pi}) \tag{6.30}$$

Considering [296], we have:

$$q(\boldsymbol{\gamma}) = \prod_{i=1}^{K}\prod_{i'=1}^{K-1} \mathcal{G}\left(c+1, d - \langle \ln\left(1 - V_{ii'}\right)\rangle\right) \tag{6.31}$$

$$q\left(\gamma_{\pi}\right) = \prod_{i=1}^{K-1} \mathcal{G}\left(\tau_{\pi 1}+1, \tau_{\pi 2} - \langle \ln\left(1 - V_{\pi i}\right)\rangle\right) \tag{6.32}$$

$$u^{\star}_{ijl} = u_{ijl} + \mathcal{U}_{ijl}, \quad v^{\star}_{ijl} = v_{ijl} - \mathcal{V}_{ijl} \tag{6.33}$$

$$\begin{aligned}\mathcal{U}_{ijl} = \sum_{p=1}^{P} \langle Z_{pij}\rangle \bar{\alpha}_{ijl}[\Psi(\sum_{d=1}^{D}\bar{\alpha}_{ijd}) - \Psi(\bar{\alpha}_{ijl}) \\ + \sum_{d=1, d\neq l}^{D} \Psi'(\sum_{d=1}^{D}\bar{\alpha}_{ijd})\bar{\alpha}_{ijd}(\langle \ln(\alpha_{ijd})\rangle - \ln(\bar{\alpha}_{ijd}))]\end{aligned} \tag{6.34}$$

$$\mathcal{V}_{ijl} = \sum_{p=1}^{P} \langle Z_{pjd}\rangle \left[\ln x_{pl} - \ln(1 - x_{pl}) - \ln\left[1 + \sum_{d=1}^{D} \frac{x_d}{(1 - x_{pl})}\right]\right] \tag{6.35}$$

$\psi(.)$ and $\psi'(.)$ in the above equations represent the digamma and trigamma functions. The value of $Z_{pij} = 1$ if X_{pt} belongs to state i and mixture component j and zero, otherwise. Thus, $\langle Z_{pij}\rangle = \sum_{t=1}^{T} \phi^{C}_{pijt} = p(s = i, m = j \mid X)$ and we compute responsibilities through a simple forward-backward procedure [297].

$$\pi^{\star}_{i} \triangleq \exp[\langle \ln(\pi_i)\rangle_{q(\pi)}] \tag{6.36}$$

6.4 Experimental Results

We tested our algorithm in human activity recognition (HAR). Providing information and discovering knowledge about individuals' physical activities is one of the most attractive and important topics in numerous fields of science and technology. Human activity recognition using various types of devices and sensor networks is broadly used in a vast range of applications such as health, athletics, and senior monitoring, rehabilitation, improving well-being, discovering patterns, and detecting activities for security. Several scientists have focused on this complex subject, however, there are lots of aspects to be addressed. In this application, data are collected by wearable, object, and ambient sensors. For instance, in medicine, caregivers could monitor and recognize the activities of patients who are suffering from morbid obesity, diabetes, dementia, or other mental disorders. This helps the healthcare system by preventing undesirable consequences based on predicting abnormal activities. Due to the sensitivity of domains in which HAR could be used, we tested our algorithm on this application as there are still issues for investigation in realistic conditions. We chose a real dataset, called opportunity [298,299], in which information was collected with three types of sensors including external and wearable sensors. Figures 6.3 and 6.4 show the set up and some types of sensors. Some of these sensors were fixed in points of interest and the others were attached to volunteer users.

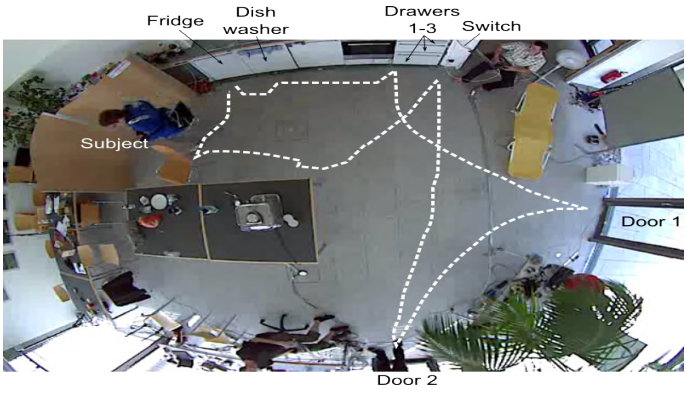


Figure 6.3: Platform and sensor setup

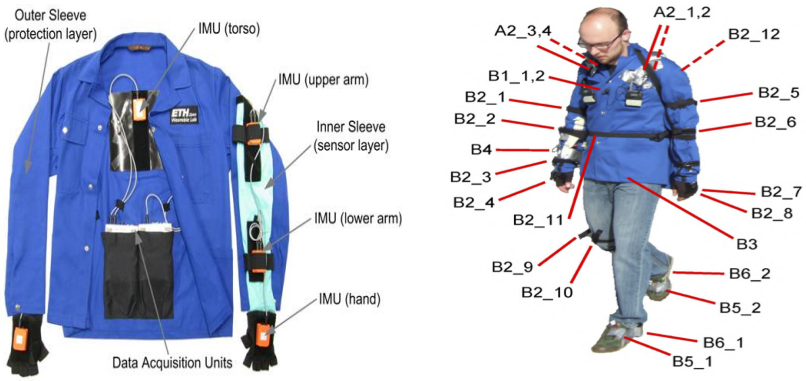


Figure 6.4: Wearable sensors.

This system was able to recognize activities of different levels as shown in Figure 6.5. The detailed information about sensors are as follows:
- Body-worn sensors: 7 inertial measurement units (IMUs), 12 3D acceleration sensors, 4 3D coordinates from a localization system.
- Object sensors: 12 objects are instrumented with wireless sensors measuring 3D acceleration and 2D rate of turn.
- Ambient sensors: 13 switches and 8 3D acceleration sensors in kitchen appliances and furniture.
The experiment is based on data collected from 4 users and 6 runs per users including 5 Activity of Daily Living (ADL) and one "drill" run. ADL is associated with a very natural manner of daily activities and in a drill, individuals execute a scripted sequence of activities.

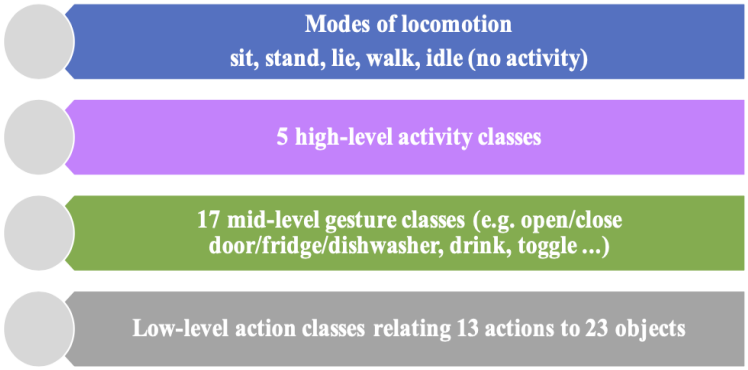


Figure 6.5: Different levels of activities.

We consider here the mode of collection for four actions: standing, walking, lying, and sitting. Also, we focused on the first individual and her/his 2 runs of activities (first and third) and test our algorithms on them.

First Individual, First Run of Activities:

This dataset includes 4 activities of the first individual and has 108 features. By analyzing data, we faced some challenges while testing our proposed algorithm on this dataset. We summarize the issues and solutions as follows:
1. Oversampling to handle unbalanced data: As it is shown in Figure 6.6, the number of instances in each cluster are very different and standing, walking, lying, and sitting have 59.7%, 17.4%, 19.9%, and 3% of share, respectively. It's worth noting that such inequality in the distribution of observations per class causes a frequency bias and our model may place more emphasis on learning from instances with more common occurrence. We tackled this issue with the help of Synthetic Minority Over-sampling Technique (SMOTE). In this approach, we generate new data points by interpolating between instances in the original dataset. So, we achieved having a balanced dataset with 22380 instances in each cluster as shown in Figure 6.7.

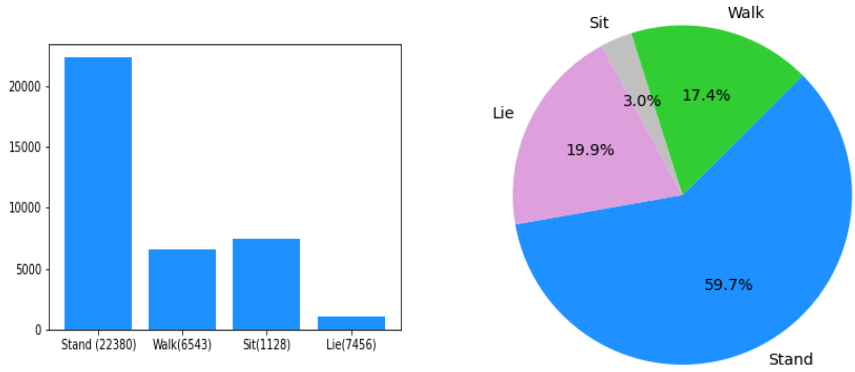


Figure 6.6: Bar and pie chart of HAR dataset 1.

Figure 6.7: Oversampling results with SMOTE.

2. Feature scaling via normalization to handle various ranges of features: The second issue that we faced was a broad range of features in the dataset. We plotted some of the features in Figure 6.8 to support our idea through visualization. These box plots indicate that the minimum and maximum values, as well as distribution of features, are so diverse. Also, Figure 6.9 illustrates some examples of feature distribution vs. activity labels. The solution to tackle this problem is normalization or Min-Max scaling. This technique shifts and re-scale values in such a way that their ranges will end up between 0 and 1. To do this, we use the following formula:

$$\vec{X} = \frac{\vec{X} - \vec{X}_{min}}{\vec{X}_{max} - \vec{X}_{min}} \tag{6.37}$$

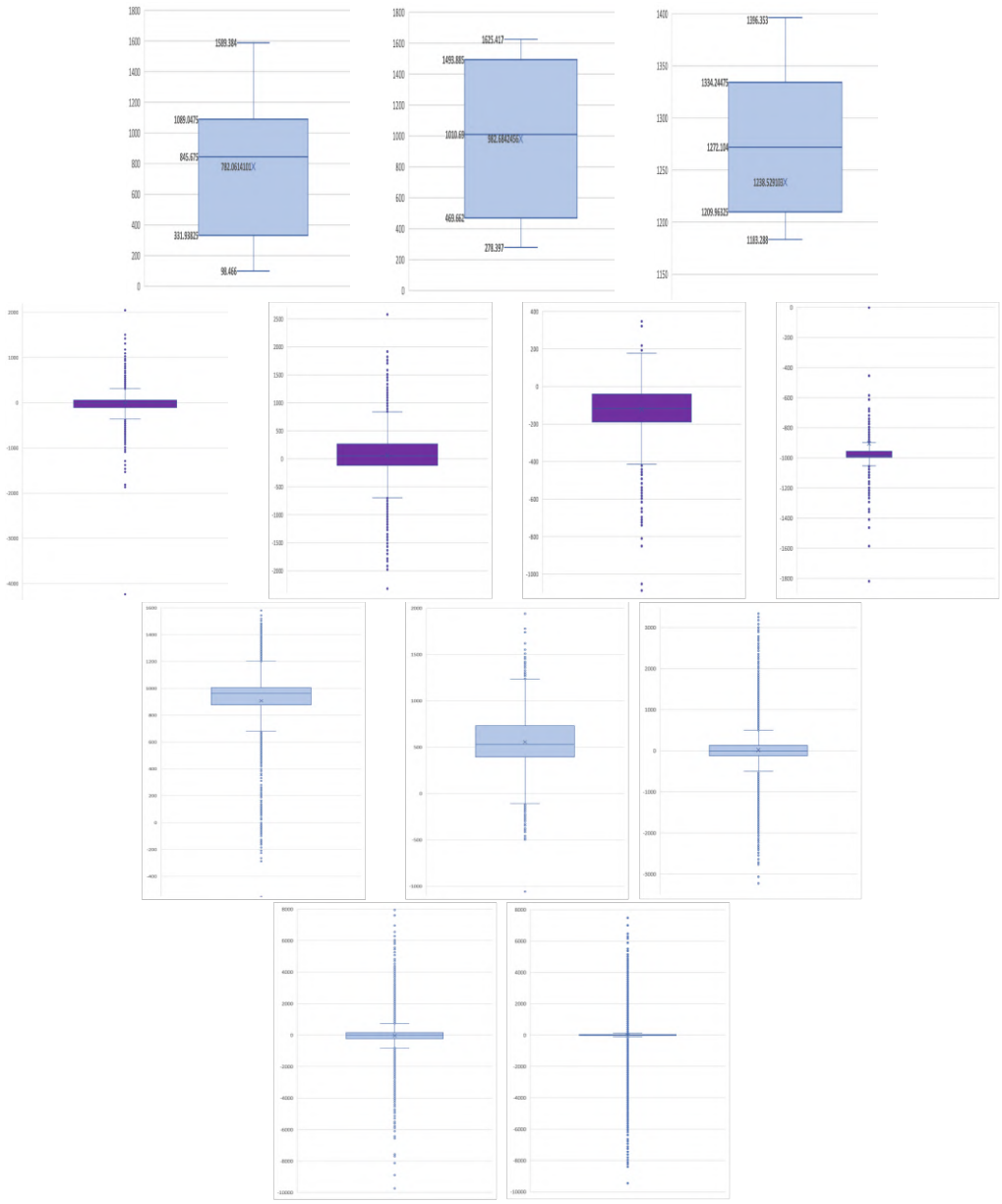


Figure 6.8: Examples of different ranges of features.

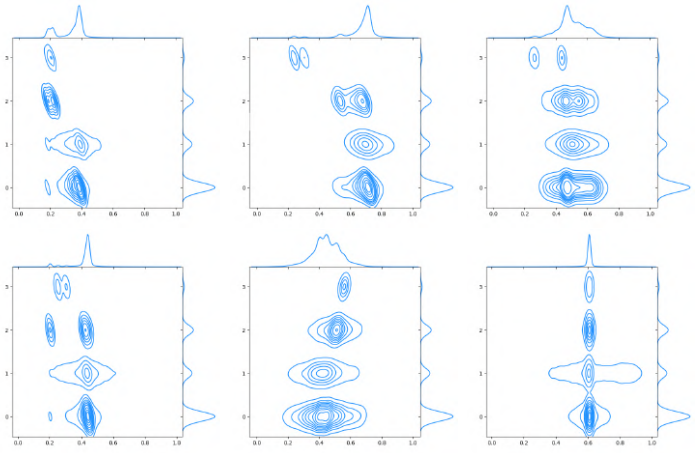


Figure 6.9: Feature distribution vs. labels.

2. Replacing missing values with the median of each feature: Similar to lots of cases while dealing with real-world applications, this dataset includes missing values. Table 6.1 indicates the number of missing values and their associated columns. As it was shown in Figure 6.8, some features have outliers. Thus, our strategy in missing value imputation and minimizing the effect of outliers is replacing them with the median of each feature.

Table 6.1: Number of Nan in each column

Column	Numbers of Nan in each column
1, 2, 3	454
4, 5, 6, 10, 11, 12 , 28, 29, 30	20
13, 14, 15	92
19, 20, 21	1681
22, 23, 24	311
34, 35, 36	37507

3. Dimensionality reduction: This dataset has 108 attributes. Figure 6.10 illustrates correlation matrix of its features. As a part of preprocessing step, we reduced the number of attributes while saving as much of the variation in the dataset as possible. This helps us to prevent some issues such as reducing computational time, increasing the overall model performance,

avoiding the curse of dimensionality, reducing the chance of over-fitting, decreasing the probability of multicollinearity and high correlation among features, and removing noise by keeping just the most important attributes. In our experiments, we applied Principal Component Analysis (PCA) to reduce dimensions.

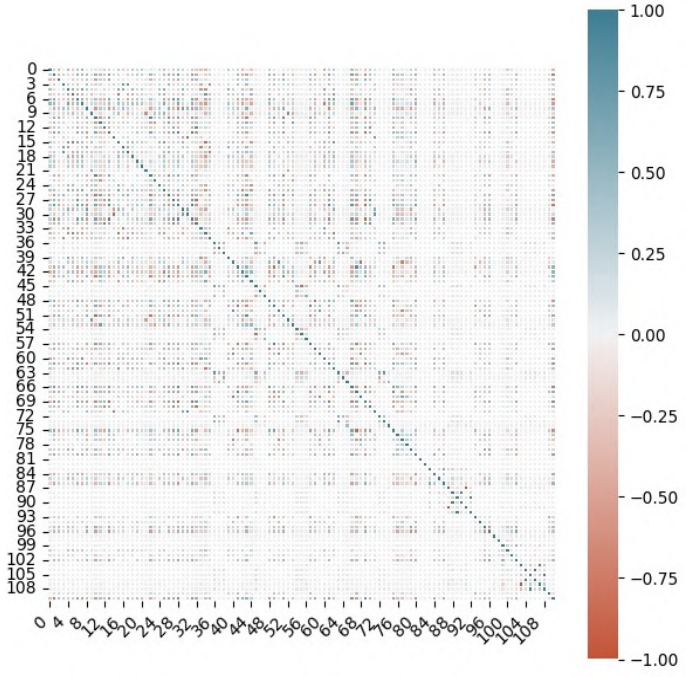


Figure 6.10: Correlation matrix.

After solving above-mentioned issues, we tested our algorithm on this dataset. To assess the model performance, we used four following criteria:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{\text{Total number of observations}} \\
Precision &= \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \\
F1 - score &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned}
\tag{6.38}
$$

TP, TN, FP and FN represent the total number of true positives, true negatives, false positives, and false negatives, respectively. Table 6.2 illustrated

the evaluation results and comparing our proposed model with similar alternatives. As it is shown, MB-HDP-HMM-VR outperforms other models by 88.33%, 88.34%, 88.33 %, 88.34% of accuracy, precision, recall and F1-score, respectively.

Table 6.2: Model performance evaluation results.

Method	Accuracy	Precision	Recall	F1-score
MB-HDP-HMM-VR	88.33	88.34	88.33	88.34
MB-HMM-VR	86.72	86.75	86.76	86.76
GMM-HMM	85.67	85.75	85.71	85.72

First Individual, Second Run of Activities:

This dataset has 25305 observations, including 10379, 6029, 7603, 1294 instances for standing, walking, sitting, lying, respectively. As illustrated in Figure 6.11, we have the same issue of unbalancing that we had in the previous dataset. We solve this problem with SMOTE and get a balanced dataset as shown in Figure 6.12.

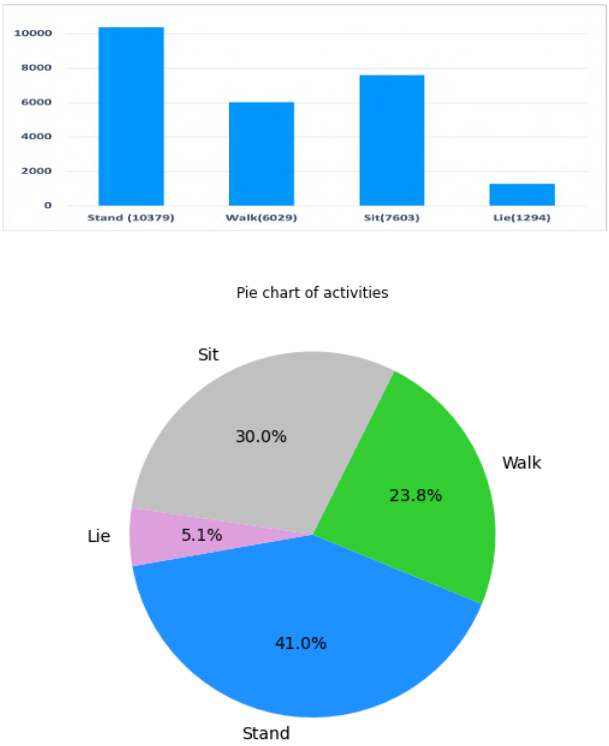


Figure 6.11: Bar and pie chart of HAR dataset 2.

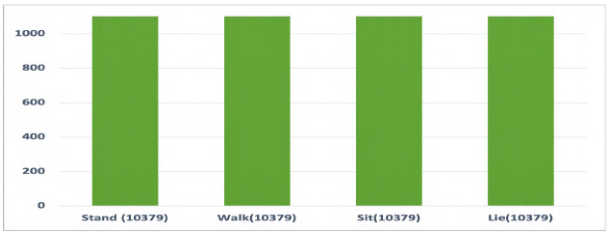


Figure 6.12: Oversampling results with SMOTE.

Moreover, we need normalization as the ranges of attributes are broadly different. Figures 6.13 and 6.14 demonstrate characteristics of some of features.

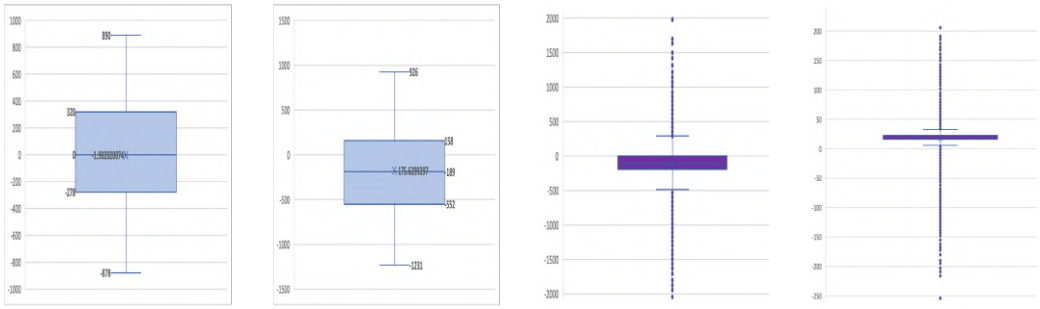


Figure 6.13: Examples of different ranges of features.

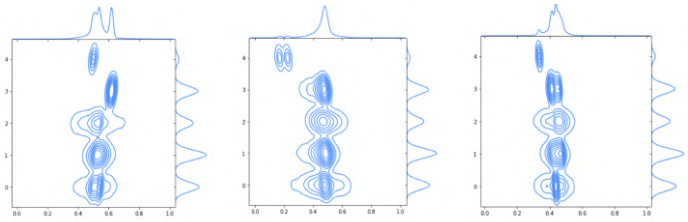


Figure 6.14: Feature distribution vs. labels

The next challenge is replacing missing values. In Figure 6.15, we show number of missing values for the attributes. We take the same strategy as the previous case and replace them with the median of attributes.

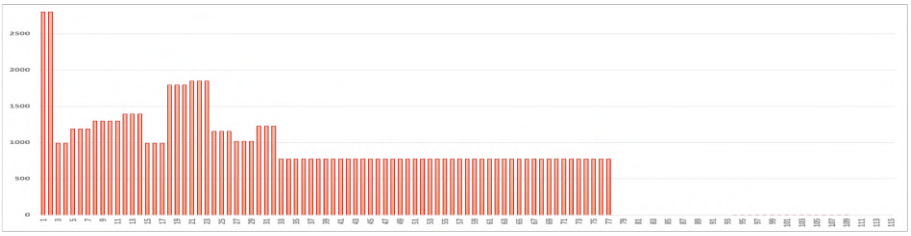


Figure 6.15: Number of missing values in each feature

To make sure that having high dimensionality won't affect model performance and to avoid potential issues that we discussed previously, we use PCA to reduce features. The correlation matrix of dataset is demonstrated in Figure 6.16.

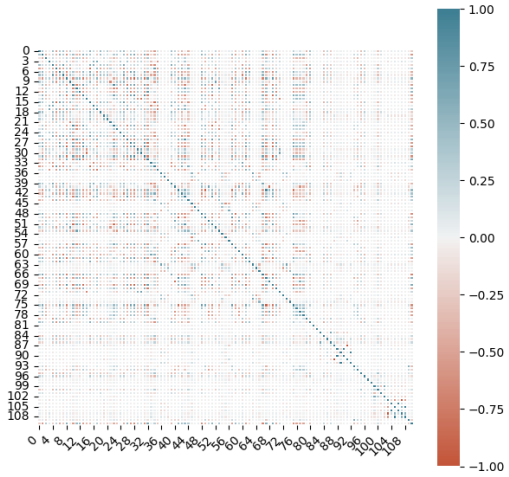


Figure 6.16: Correlation matrix

Table 6.3 illustrated the evaluation results of comparing our proposed model with similar alternatives. MB-HDP-HMM-VR has improved robustness with 86.43, 86.66, 88.43, 86.55 percentage of accuracy, precision, recall and F1-score, respectively.

Table 6.3: Model performance evaluation results

Method	Accuracy	Precision	Recall	F1-score
MB-HDP-HMM-VR	86.43	86.66	88.43	86.55
MB-HMM-VR	84.88	84.87	84.88	84.87
GMM-HMM	84.37	84.39	84.37	84.38

In Figure 6.17, we compare the results of testing our model on two datasets. We have better results in the first dataset considering these graphs. One of the causes could be having more data points in the first dataset as its size is twice larger than the second dataset (22380 vs. 10379 in each cluster).

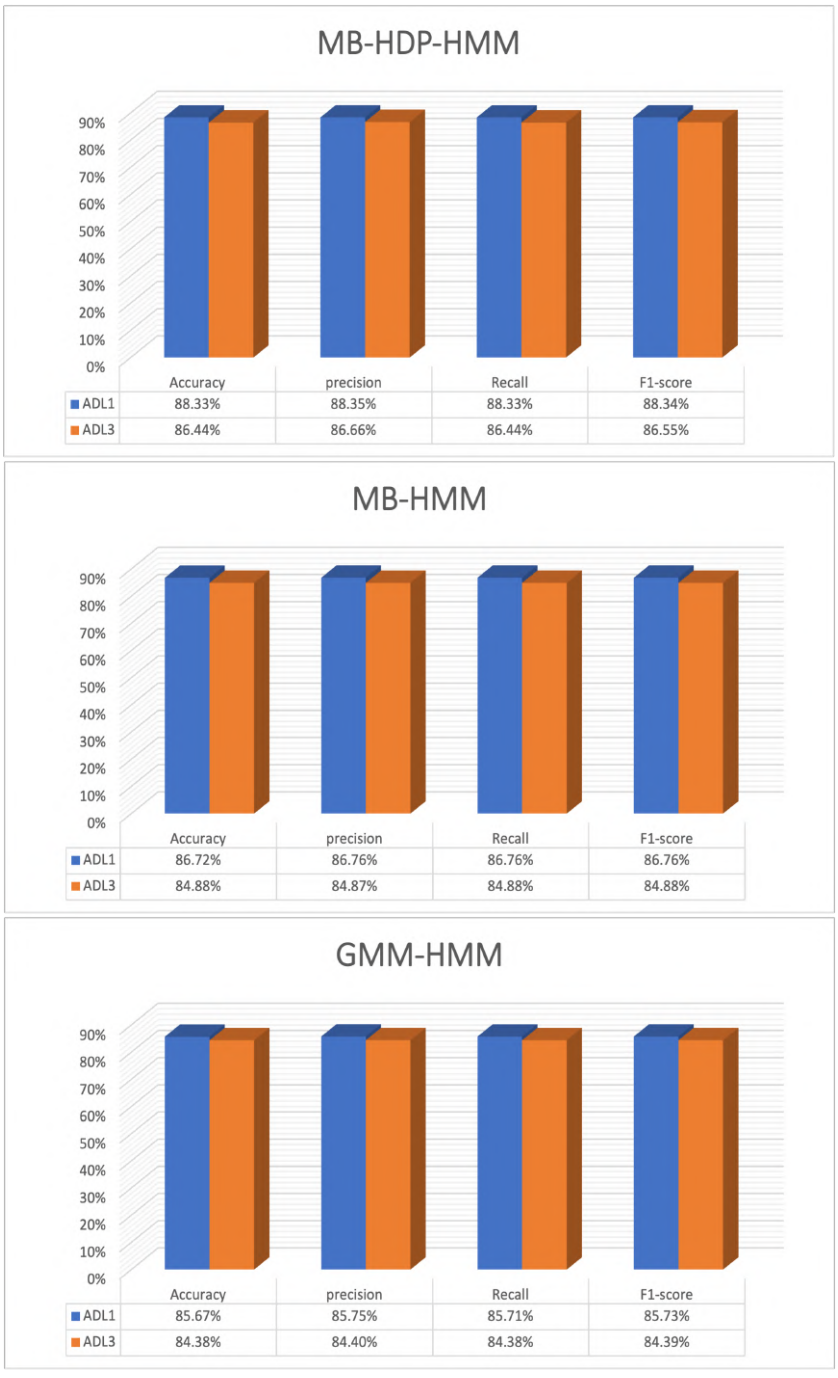


Figure 6.17: Results comparison.

6.5 Conclusion

In this paper, we proposed multivariate Beta-based hierarchical Dirichlet process hidden Markov models as a new extension of HMMs and applied it to two real datasets. The nonparametric structure of this model assists in handling issues such as defining the number of states. Another motivation to work on this novel algorithm was that we can not generalize the assumption of Gaussianity in all cases. Over the past decades, other alternative distributions have been applied to numerous real-world datasets. One of the proper choices is multivariate Beta distribution which has demonstrated good potential and flexibility in fitting data. By changing its shape parameter, we could model data with various shapes such as symmetric, asymmetric, skewed ones. In our model, we assumed that emission probability distributions follow multivariate Beta mixture models. This modification may result in having better outputs compared to the conventional cases where we consider GMM-based HMM. To learn the model, we applied variational inference which is slightly faster than fully Bayesian inference and more precise compared to deterministic methods. This promising strategy is successful in various domains. Finally, we evaluated our model on two real datasets, and considering the outcomes, we could infer that our proposed model demonstrates more robustness. In future steps, we can focus on feature selection and integrate it into our model.

Chapter 7

Conclusion

In this thesis, we proposed several unsupervised methods and applied them to medical cases. We formulated powerful alternatives to widely used algorithms such as Gaussian mixture models and GMM-based hidden Markov models. We developed our models based on a new distribution, multivariate Beta distribution. Our motivation was the fact that the assumption of Gaussianity is not valid for many of data sets in different fields of sciences and real-world applications. Thus, we chose a more flexible alternative to be able to model symmetric, asymmetric, and skewed data. We proposed first the finite case. Later on, we borrowed the idea from elegant non-parametric models and extended the finite case to Dirichlet process of multivariate Beta distribution. To increase the power of the model, we improved this version of our non-parametric model by using hierarchical Dirichlet process of multivariate Beta distribution with two levels of hierarchy. This modification enabled our model to capture the hidden pattern of data. Also, it could automatically define the proper number of clusters which is one of the main concerns while we use mixture models. We paired our model with variational inference as an efficient and accurate learning technique compared to conventional methods such as deterministic and fully Bayesian inferences. In finite mixture model, variational inference allowed us to define the model complexity and estimate parameters simultaneously. Also, we applied another approximation learning, expectation propagation which is relatively fast and more accurate compared to variational learning. In real world, data have a sequential nature. This motivated us to focus on hidden Markov models. We proposed a modification in the conventional model where emission prob-

abilities have been traditionally assumed to follow Gaussian mixture models. In our novel method, we assumed that emission probabilities are raised from multivariate Beta mixture models. Also, we assumed a hierarchical structure inside our mixture model and we called our newly proposed model, multivariate Beta-based hierarchical Dirichlet process hidden Markov model. We learned this new model with variational learning too.
In all chapters, we evaluated our models on medical images such as pathological samples of colon, bone, and oral tissues to differentiate between malignant and benign cases. Also, we applied our models to other applications such as malaria detection as well as skin, and white blood cell analysis. We conducted our research on medical signals to analyze sentiments based on EEG. Another interesting in this field was human activity recognition based on sensor data.
We compared our models with other similar choices and measured the performance of our models. Based on reported results, our proposed models provide better outcomes compared to other alternatives and have the potential to be considered as second-opinion systems.
To extend and improve our research, we are planning to work on the following topics in future steps:

- Integrating feature selection into our proposed models.
- Working on other distributions which could be considered as capable alternatives.
- Extending our models to the infinite mixture of infinite multivariate Beta mixture.
- Working on other medical applications with sequential data such as chemotherapy treatments.
- Integrating our proposed models into other models to develop novel hybrid algorithms.

Appendix A

Proof of Equations of Chapter 2

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

The variational solution $Q_s(\Theta_s)$ is expressed by:

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

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

2. Proof of equation (2.16) variational solution of $Q(\mathcal{Z})$:

$$\begin{aligned} \ln Q(Z_{ij}) = Z_{ij} \Bigg[& \ln \pi_j + R_j + \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} - 1\big) \ln x_{il} \\ & - \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} + 1\big) \ln(1 - x_{il}) \\ & - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})} \Big] \Bigg] + \text{const} \end{aligned} \tag{A.2}$$

where,

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

$$\big\langle \alpha_{jl} \big\rangle = \frac{u_{jl}}{\nu_{jl}} \tag{A.4}$$

As R_j is intractable and has not a closed form and standard variational inference can be applied indirectly. Thus, we approximate the lower bound to obtain a closed-form expression by the second-order Taylor series expansion. The function R_j is approximated about $\vec{\alpha}$. $\tilde{R}_j$ and $(\overline{\alpha}_{j1}, \ldots, \overline{\alpha}_{jD})$ are notations for approximation of R_j and $\vec{\alpha}$, respectively. The approximation of R_j is proved in [64] and after replacing it by $\tilde{R}_j$, optimization of (A.2) is tractable. So, the optimal solution for $\mathcal{Z}$ can be derived by:

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

$$\begin{aligned} \ln \tilde{r}_{ij} =& \ln \pi_j + \tilde{R} + \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} - 1\big) \ln x_{il} \\ &- \sum_{l=1}^{D} \big(\overline{\alpha}_{jl} + 1\big) \ln(1 - x_{il}) - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{l=1}^{D} \frac{x_{il}}{(1 - x_{il})}\Big] + \text{const} \end{aligned} \tag{A.6}$$

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

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

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

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

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

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

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

2. Proof of equation (2.17): variational solution of $Q(\vec{\alpha})$

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

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

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

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

where,

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

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

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

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

$$
\begin{aligned}
\ln Q\big(\alpha_{js}\big) = & \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \overline{\alpha}_{js} \ln \alpha_{js} \Bigg[\psi\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) - \psi\big(\overline{\alpha}_{js}\big) \\
& + \sum_{s \neq l}^{D} \overline{\alpha}_{js} \psi'\Bigg(\sum_{l=1}^{D} \overline{\alpha}_{jl} \Bigg) \big(\big\langle \ln \alpha_{jl} \big\rangle - \ln \overline{\alpha}_{jl} \big) \Bigg] \\
& + \alpha_{js} \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln x_{is} - \ln(1 - x_{is}) \\
& - \ln \Big[1 + \sum_{l=1}^{D} \frac{x_{is}}{(1 - x_{is})} \Big] \Bigg] \\
& + \big(u_{jl} - 1 \big) \ln \alpha_{jl} - \nu_{jl} \alpha_{jl} + \text{const} \\
& = \ln \alpha_{js} \big(u_{js} + \varphi_{js} - 1 \big) - \alpha_{js} \big(\nu_{js} - \vartheta_{js} \big) + \text{const}
\end{aligned}
\tag{A.15}
$$

where,

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

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

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

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

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

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

3. Calculation of $\tilde{R}_j$ for equations (2.19) and (2.32)

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

4. Lower bound $\mathcal{L}\big(Q\big)$ of (2.28)

$$\mathcal{L}\big(Q\big) = \sum_{\mathcal{Z}} \int Q\big(\mathcal{Z}, \vec{\alpha}\big) \ln \left(\frac{p\big(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}\big)}{Q\big(\mathcal{Z}, \vec{\alpha}\big)} \right) \mathrm{d}\vec{\alpha} \tag{A.21}$$

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

Appendix B

Proof of Equations of Chapter 4

1. Hyperparameters of equations (4.30) to (4.35):

ρ_{jit} as the hyperparameter of $Q(\vec{Z})$ is found by:

$$\rho_{jit} = \frac{\exp(\tilde{\rho}_{jit})}{\sum_{f=1}^{T} exp(\tilde{\rho}_{jit})} \tag{B.1}$$

$$\tilde{\rho}_{jit} = \sum_{k=1}^{K} \langle W_{jtk} \rangle \Bigg[\tilde{R}_k + \sum_{d=1}^{D} \big(\overline{\alpha}_{kd} - 1\big) \ln Y_{jid} - \sum_{d=1}^{D} \big(\overline{\alpha}_{kd} + 1\big) \ln(1 - Y_{jid}) \tag{B.2}$$
$$- \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{d=1}^{D} \frac{Y_{jid}}{(1 - Y_{jid})}\Big] + \langle \ln \pi'_{jt} \rangle + \sum_{s=1}^{t-1} \langle \ln(1 - \pi'_{js}) \rangle \Bigg]$$

$\tilde{R}_k$ following [192] is calculated by:

$$\tilde{R}_j = \ln \frac{\Gamma\big(\sum_{d=1}^{D} \overline{\alpha}_{jd}\big)}{\prod_{d=1}^{D} \Gamma\big(\overline{\alpha}_{jd}\big)} + \tag{B.3}$$
$$\sum_{d=1}^{D} \overline{\alpha}_{jd} \left[\Psi\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) - \Psi\big(\overline{\alpha}_{jd}\big) \right] \times \Big[\big\langle \ln \alpha_{jd} \big\rangle - \ln \overline{\alpha}_{jd} \Big]$$
$$+ \frac{1}{2} \sum_{d=1}^{D} \overline{\alpha}_{jd}^{2} \left[\Psi'\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) - \Psi'\big(\overline{\alpha}_{jd}\big) \right] \times \Big\langle \big(\ln \alpha_{jd} - \ln \overline{\alpha}_{jd} \big)^2 \Big\rangle + \frac{1}{2} \sum_{a=1}^{D}$$
$$\sum_{b=1, a \neq b}^{D} \overline{\alpha}_{ja}\, \overline{\alpha}_{jb} \left[\Psi'\left(\sum_{d=1}^{D} \overline{\alpha}_{jd} \right) \times \Big(\big\langle \ln \overline{\alpha}_{ja} \big\rangle - \ln \overline{\alpha}_{ja} \Big) \times \Big(\big\langle \ln \overline{\alpha}_{jb} \big\rangle - \ln \overline{\alpha}_{jb} \Big) \right]$$

$\Psi(.)$ and $\Psi'(.)$ are the digamma and trigamma functions, respectively. Similarly, ϑ_{jtk} of the factor $Q(W)$ is defined by:

$$\vartheta_{jtk} = \frac{\exp(\tilde{\vartheta}_{jtk})}{\sum_{f=1}^{K} exp(\tilde{\vartheta}_{jtf})} \tag{B.4}$$

$$\tilde{\vartheta}_{jtk} = \sum_{i=1}^{N} \big\langle Z_{jit} \big\rangle \left[\tilde{R}_k + \sum_{d=1}^{D} \big(\overline{\alpha}_{kd} - 1 \big) \ln Y_{jid} - \sum_{d=1}^{D} \big(\overline{\alpha}_{kd} + 1 \big) \ln(1 - Y_{jid}) \right. \tag{B.5}$$
$$\left. - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{d=1}^{D} \frac{Y_{jid}}{(1 - Y_{jid})} \Big] + \langle \ln \psi'_k \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \psi'_s) \rangle \right]$$

a_{jt} and b_{jt} as the hyperparameters of the factor $Q(\pi')$ are:

$$a^*_{jt} = 1 + \sum_{i=1}^{N} \big\langle Z_{jit} \big\rangle, \quad b^*_{jt} = \lambda_{jt} + \sum_{i=1}^{N} \sum_{s=t+1}^{T} \langle Z_{jis} \rangle \tag{B.6}$$

c_k and d_k as the hyperparameters of the factor $Q(\psi')$ are calculated by:

$$c^*_k = 1 + \sum_{j=1}^{M} \sum_{t=1}^{T} \big\langle W_{jtk} \big\rangle, \quad d^*_k = \gamma_k + \sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{s=k+1}^{K} \langle W_{jts} \rangle \tag{B.7}$$

The hyperparameters u^*_{kd} and ν_{kd} of the factor $Q(\alpha)$ are updated as follows:

$$u^*_{kd} = u_{kd} + \sum_{j=1}^{M}\sum_{t=1}^{T}\langle W_{jtk}\rangle \sum_{i=1}^{N}\left\langle Z_{jit}\right\rangle \overline{\alpha}_{kd}\times \tag{B.8}$$
$$\left[\Psi\left(\sum_{s=1}^{D}\overline{\alpha}_{ks}\right) - \Psi\big(\overline{\alpha}_{kd}\big) + \sum_{s\neq l}^{D}\overline{\alpha}_{ks}\Psi'\left(\sum_{s=1}^{D}\overline{\alpha}_{ks}\right)\times\Big(\left\langle \ln\alpha_{ks}\right\rangle - \ln\overline{\alpha}_{ks}\Big)\right]$$

$$\nu^*_{kd} = \nu_{kd} - \sum_{j=1}^{M}\sum_{t=1}^{T}\langle W_{jtk}\rangle \sum_{i=1}^{N}\left\langle Z_{jit}\right\rangle \times \left[\ln Y_{jid} - \ln(1-Y_{jid}) - \right. \tag{B.9}$$
$$\left.\ln\left[1+\sum_{d=1}^{D}\frac{Y_{jid}}{(1-Y_{jid})}\right]\right]$$

The expected values of above mentioned equations are given by:

$$\bar{\alpha}_{kd} = \langle\alpha_{kd}\rangle = \frac{u^*_{kd}}{v^*_{kd}}, \tag{B.10}$$

$$\left\langle Z_{jit}\right\rangle = \rho_{jit}, \quad \left\langle W_{jtk}\right\rangle = \vartheta_{jtk} \tag{B.11}$$

$$\left\langle \ln\pi'_{jt}\right\rangle = \Psi(a_{jt}) - \Psi(a_{jt}+b_{jt}), \quad \left\langle \ln(1-\pi'_{jt})\right\rangle = \Psi(b_{jt}) - \Psi(a_{jt}+b_{jt}) \tag{B.12}$$

$$\left\langle \ln\psi'_{k}\right\rangle = \Psi(c_{k}) - \Psi(c_{k}+d_{k}), \quad \left\langle \ln(1-\psi'_{k})\right\rangle = \Psi(d_{k}) - \Psi(c_{k}+d_{k}) \tag{B.13}$$

$$\langle\ln\alpha_{kd}\rangle = \Psi(u^*_{kd}) - \ln v^*_{kd}, \tag{B.14}$$
$$\left\langle\big(\ln\alpha_{kd} - \ln\overline{\alpha}_{kd}\big)^2\right\rangle = \left[\Psi(u^*_{kd}) - \ln u^*_{kd}\right]^2 + \Psi'(u^*_{kd})$$

2. Hyperparameters of equation (4.37)**:**

$$\rho_{jtr} = \frac{\exp(\tilde{\rho}_{jtr})}{\sum_{f=1}^{T} \exp(\tilde{\rho}_{jtr})} \tag{B.15}$$

$$\begin{aligned}\tilde{\rho}_{jtr} = \sum_{k=1}^{K} \left\langle W_{jtk}^{(r-1)} \right\rangle \Bigg[\tilde{R}_k{}^{(r-1)} + \sum_{d=1}^{D} \left(\overline{\alpha}_{kd}{}^{(r-1)} - 1 \right) \ln X_{jrl} \\ - \sum_{d=1}^{D} \left(\overline{\alpha}_{kd}{}^{(r-1)} + 1 \right) \ln(1 - X_{jrl}) - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{d=1}^{D} \frac{X_{jrl}}{(1 - X_{jrl})} \Big] + \\ \langle \ln \pi_{jt}^{\prime(r-1)} \rangle + \sum_{s=1}^{t-1} \langle \ln(1 - \pi_{js}^{\prime(r-1)}) \rangle \Bigg]\end{aligned} \tag{B.16}$$

$\tilde{R}_j$ is calculated by equation (B.3).

3. Equation (4.40)**:**

$$\begin{aligned}\tilde{\vartheta}_{jtk} = N \rho_{jtr} \Bigg[\tilde{R}_k{}^{(r-1)} + \sum_{d=1}^{D} \left(\overline{\alpha}_{kd}{}^{(r-1)} - 1 \right) \ln X_{jrl} \\ - \sum_{d=1}^{D} \left(\overline{\alpha}_{kd}{}^{(r-1)} + 1 \right) \ln(1 - X_{jrl}) \\ - \mid \overline{\alpha}_j \mid \ln \Big[1 + \sum_{d=1}^{D} \frac{X_{jrl}}{(1 - X_{jrl})} \Big] + \langle \ln \psi_k^{\prime(r-1)} \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \psi_k^{\prime(r-1)}) \rangle \Bigg]\end{aligned} \tag{B.17}$$

4. Hyperparameters of equations (4.41) **to** (4.43)**:**

$$a_{jt}^{(r)} = a_{jt}^{(r-1)} + \xi_r \Delta a_{jt}^{(r)}, \quad b_{jt}^{(r)} = b_{jt}^{(r-1)} + \xi_r \Delta b_{jt}^{(r)} \tag{B.18}$$

$$c_k^{(r)} = c_k^{(r-1)} + \xi_r \Delta c_k^{(r)}, \quad d_k^{(r)} = d_k^{(r-1)} + \xi_r \Delta d_k^{(r)} \tag{B.19}$$

$$u_{kd}^{*(r)} = u_{kd}^{*(r-1)} + \xi_r \Delta u_{kd}^{*(r)}, \quad v_{kd}^{*(r)} = v_{kd}^{*(r-1)} + \xi_r \Delta v_{kd}^{*(r)} \tag{B.20}$$

$$\Delta a_{jt}^{(r)} = 1 + N\rho_{jtr} - a_{jt}^{(r-1)}, \quad \Delta b_{jt}^{(r)} = \lambda_{jt} + N \sum_{s=t+1}^{T} \rho_{jsr} - b_{jt}^{(r-1)} \tag{B.21}$$

$$\Delta c_k^{(r)} = 1 + \sum_{j=1}^{K} \sum_{t=1}^{T} \vartheta_{jtk}^{(r)} - c_k^{(r-1)}, \quad \Delta d_k^{(r)} = \gamma_k + \sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{m=k+1}^{K} \vartheta_{jtk}^{(r)} - d_k^{(r-1)} \tag{B.22}$$

$$\Delta u_{kd}^{*(t)} = u_{kd} + N \sum_{j=1}^{M} \sum_{t=1}^{T} \vartheta_{jtk}^{(r)} \rho_{jtr} \bar{\alpha}_{kd}^{(r-1)} \big[\Psi(\sum_{s=1}^{D} \bar{\alpha}_{ks}^{(r-1)}) - \Psi(\bar{\alpha}_{kd}^{(r-1)}) +) \tag{B.23}$$

$$\sum_{s \neq l}^{D} \bar{\alpha}_{ks}^{(r-1)} \Psi'(\sum_{s=1}^{D} \bar{\alpha}_{ks}^{(r-1)}) (\langle \ln \alpha_{ks}^{(r-1)} \rangle - \ln \bar{\alpha}_{ks}^{(r-1)} - \ln \bar{\alpha}_{ks}^{(r-1)} \big] - u_{kd}^{*(t-1)}$$

$$\Delta v_{kd}^{*(t)} = v_{kd} - N \sum_{j=1}^{M} \sum_{t=1}^{T} \vartheta_{jtk}^{(r)} \rho_{jtr} \Bigg[\ln Y_{jid} - \ln(1 - Y_{jid}) - \tag{B.24}$$

$$\ln \Big[1 + \sum_{d=1}^{D} \frac{Y_{jid}}{(1 - Y_{jid})} \Big] \Bigg] - v_{kd}^{*(t-1)}$$

List of References

[1] Mamoru Tokunaga, Tomoaki Matsumura, Rino Nankinzan, Takuto Suzuki, Hirotaka Oura, Tatsuya Kaneko, Mai Fujie, Shun Hirai, Ryota Saiki, Naoki Akizue, et al. Computer-aided diagnosis system using only white-light endoscopy for the prediction of invasion depth in colorectal cancer. *Gastrointestinal Endoscopy*, 93(3):647–653, 2021.

[2] Chieh Sian Koo, Dmitrii Dolgunov, and Calvin Jianyi Koh. Key tips for using computer-aided diagnosis in colonoscopy–observations from two different platforms. *Endoscopy*, 2021.

[3] Clayton R Pereira, Danilo R Pereira, Silke AT Weber, Christian Hook, Victor Hugo C De Albuquerque, and Joao P Papa. A survey on computer-assisted parkinson's disease diagnosis. *Artificial intelligence in medicine*, 95:48–63, 2019.

[4] Quirine EW van der Zander, Ramon M Schreuder, Roger Fonollà, Thom Scheeve, Fons van der Sommen, Bjorn Winkens, Patrick Aepli, Bu'Hussain Hayee, Andreas B Pischel, Milan Stefanovic, et al. Optical diagnosis of colorectal polyp images using a newly developed computer-aided diagnosis system (cadx) compared with intuitive optical diagnosis. *Endoscopy*, 53(12):1219–1226, 2021.

[5] Toshio Uraoka, Shinji Tanaka, Yutaka Saito, Takayuki Matsumoto, Shiko Kuribayashi, Keisuke Hori, and Hisao Tajiri. Computer-assisted detection of diminutive and small colon polyps by colonoscopy using an extra-wide-area-view colonoscope. *Endoscopy*, 53(03):E102–E103, 2021.

[6] Yuki Okamoto, Shigeto Yoshida, Seiji Izakura, Daisuke Katayama, Ryuichi Michida, Tetsushi Koide, Toru Tamaki, Yuki Kamigaichi, Hirosato Tamari, Yasutsugu Shimohara, et al. Development of multiclass computer-aided diagnostic systems using the nice/jnet classifications for colorectal lesions. *Journal of gastroenterology and hepatology*, 37(1):104–110, 2022.

[7] David J Winkel, Angela Tong, Bin Lou, Ali Kamen, Dorin Comaniciu, Jonathan A Disselhorst, Alejandro Rodríguez-Ruiz, Henkjan Huisman, Dieter Szolar, Ivan Shabunin, et al. A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. *Investigative Radiology*, 56(10):605–613, 2021.

[8] Regine Mariette Perl, Rainer Grimmer, Tobias Hepp, and Marius Stefan Horger. Can a novel deep neural network improve the computer-aided detection of solid pulmonary nodules and the rate of false-positive findings in comparison to an established machine learning computer-aided detection? *Investigative Radiology*, 56(2):103–108, 2021.

[9] Xiaowen Liang, Yingmin Huang, Yongyi Cai, Jianyi Liao, and Zhiyi Chen. A computer-aided diagnosis system and thyroid imaging reporting and data system for dual validation of ultrasound-guided fine-needle aspiration of indeterminate thyroid nodules. *Frontiers in Oncology*, page 4037, 2021.

[10] Yuhan Yang, Yin Zhou, Chen Zhou, Xuemei Zhang, and Xuelei Ma. Mri-based computer-aided diagnostic model to predict tumor grading and clinical outcomes in patients with soft tissue sarcoma. *Journal of Magnetic Resonance Imaging*, 2022.

[11] Tsuyoshi Ozawa, Soichiro Ishihara, Mitsuhiro Fujishiro, Hiroaki Saito, Youichi Kumagai, Satoki Shichijo, Kazuharu Aoyama, and Tomohiro Tada. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointestinal endoscopy*, 89(2):416–421, 2019.

[12] Lianyan Xu, Ke Yan, Le Lu, Weihong Zhang, Xu Chen, Xiaofei Huo, and Jingjing Lu. External and internal validation of a computer as-

sisted diagnostic model for detecting multi-organ mass lesions in ct images. *Chinese Medical Sciences Journal*, 36(3):210–217, 2021.

[13] Pritesh Mehta, Michela Antonelli, Hashim U Ahmed, Mark Emberton, Shonit Punwani, and Sébastien Ourselin. Computer-aided diagnosis of prostate cancer using multiparametric mri and clinical features: A patient-level classification framework. *Medical image analysis*, 73:102153, 2021.

[14] Wendie A Berg, David Gur, Andriy I Bandos, Bronwyn Nair, Terri-Ann Gizienski, Cathy S Tyma, Gordon Abrams, Katie M Davis, Amar S Mehta, Grace Rathfon, et al. Impact of original and artificially improved artificial intelligence–based computer-aided diagnosis on breast us interpretation. *Journal of Breast Imaging*, 3(3):301–311, 2021.

[15] Paul GM Knoops, Athanasios Papaioannou, Alessandro Borghi, Richard WF Breakey, Alexander T Wilson, Owase Jeelani, Stefanos Zafeiriou, Derek Steinbacher, Bonnie L Padwa, David J Dunaway, et al. A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific reports*, 9(1):1–12, 2019.

[16] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

[17] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020.

[18] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

[19] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian,

Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[21] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

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

[23] Stephen Adams and Peter A Beling. A survey of feature selection methods for gaussian mixture models and hidden markov models. *Artificial Intelligence Review*, 52(3):1739–1779, 2019.

[24] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412, 2003.

[25] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Expectation propagation learning of finite multivariate beta mixture models and applications. *Neural Computing and Applications*, pages 1–11, 2022.

[26] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Batch and online variational learning of hierarchical pitman-yor mixtures of multivariate beta distributions. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 298–303. IEEE, 2021.

[27] Narges Manouchehri and Nizar 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.

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

[29] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Batch and online variational learning of hierarchical dirichlet process mixtures of

multivariate beta distributions in medical applications. *Pattern Analysis and Applications*, 24(4):1731–1744, 2021.

[30] Narges Manouchehri, Meeta Kalra, and Nizar Bouguila. Online variational inference on finite multivariate beta mixture models for medical applications. *IET Image Processing*, 15(9):1869–1882, 2021.

[31] Narges Manouchehri and Nizar Bouguila. A frequentist inference method based on finite bivariate and multivariate beta mixture models. In *Mixture Models and Applications*, pages 179–208. Springer, 2020.

[32] Narges Manouchehri, Hieu Nguyen, and Nizar Bouguila. Component splitting-based approach for multivariate beta mixture models learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019.

[33] Narges Manouchehri, Maryam Rahmanpour, Nizar Bouguila, and Wentao Fan. Learning of multivariate beta mixture models via entropy-based component splitting. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2825–2832. IEEE, 2019.

[34] Narges Manouchehri and Nizar Bouguila. A probabilistic approach based on a finite mixture model of multivariate beta distributions. In *International Conference on Enterprise Information Systems (ICEIS) (1)*, pages 373–380, 2019.

[35] Mahsa Amirkhani, Narges Manouchehri, and Nizar Bouguila. Birth-death mcmc approach for multivariate beta mixture models in medical applications. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 285–296. Springer, 2021.

[36] Mahsa Amirkhani, Narges Manouchehri, and Nizar 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.

[37] Narges Manouchehri and Nizar Bouguila. Multivariate beta-based hierarchical dirichlet process hidden markov models in medical applications. In *Hidden Markov Models and Applications*, pages 1–28. Springer, 2022.

[38] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[39] Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. A survey on deep learning for big data. *Information Fusion*, 42:146–157, 2018.

[40] LC Yan, B Yoshua, and H Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015.

[41] Guangle Yao, Tao Lei, and Jiandan Zhong. A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, 118:14–22, 2019.

[42] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[43] David Freire-Obregón, Fabio Narducci, Silvio Barra, and Modesto Castrillón-Santana. Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*, 126:86–91, 2019.

[44] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[45] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

[46] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[47] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[48] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[49] Farnaam Samadi, Gholamreza Akbarizadeh, and Hooman Kaabi. Change detection in sar images using deep belief network: a new training approach based on morphological images. *IET Image Processing*, 13(12):2255–2264, 2019.

[50] Foroogh Sharifzadeh, Gholamreza Akbarizadeh, and Yousef Seifi Kavian. Ship classification in sar images using a new hybrid cnn–mlp classifier. *Journal of the Indian Society of Remote Sensing*, 47(4):551–562, 2019.

[51] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE transactions on geoscience and remote sensing*, 56(5):2811–2821, 2018.

[52] Salman Khan, Khan Muhammad, Shahid Mumtaz, Sung Wook Baik, and Victor Hugo C de Albuquerque. Energy-efficient deep cnn for smoke detection in foggy iot environment. *IEEE Internet of Things Journal*, 6(6):9237–9245, 2019.

[53] Alex Adim Obinikpo and Burak Kantarci. Big sensed data meets deep learning for smarter health care in smart cities. *Journal of Sensor and Actuator Networks*, 6(4):26, 2017.

[54] MN Arun Kumar, MN Kumar, and HS Sheshadri. Computer aided detection of clustered microcalcification: A survey. *Current Medical Imaging*, 15(2):132–149, 2019.

[55] Volker Tresp, J Marc Overhage, Markus Bundschus, Shahrooz Rabizadeh, Peter A Fasching, and Shipeng Yu. Going digital: a survey on digitalization and large-scale data analytics in healthcare. *Proceedings of the IEEE*, 104(11):2180–2206, 2016.

[56] Jose Bernal, Kaisar Kushibar, Daniel S Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 95:64–81, 2019.

[57] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation

in pathology. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019.

[58] Paschalis Bizopoulos and Dimitrios Koutsouris. Deep learning in cardiology. *IEEE reviews in biomedical engineering*, 12:168–193, 2018.

[59] Kaushiki Roy, Debapriya Banik, Debotosh Bhattacharjee, and Mita Nasipuri. Patch-based system for classification of breast histology images using deep learning. *Computerized Medical Imaging and Graphics*, 71:90–103, 2019.

[60] Roopa B Hegde, Keerthana Prasad, Harishchandra Hebbar, and Brij Mohan Kumar Singh. Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. *Biocybernetics and Biomedical Engineering*, 39(2):382–392, 2019.

[61] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: a preliminary study. *Radiology*, 286(3):887–896, 2018.

[62] Fukui Tian, Qingyi Zhou, and Chuanchuan Yang. Gaussian mixture model-hidden markov model based nonlinear equalizer for optical fiber transmission. *Optics Express*, 28(7):9728–9737, 2020.

[63] 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.

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

[65] Nizar Bouguila and Djemel Ziou. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1716–1731, 2007.

[66] Wentao Fan and Nizar Bouguila. Variational learning of a dirichlet process of generalized dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognition*, 46(10):2754–2769, 2013.

[67] Wentao Fan and Nizar Bouguila. Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE transactions on neural networks and learning systems*, 24(11):1850–1862, 2013.

[68] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

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

[70] 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.

[71] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. *arXiv preprint arXiv:1301.6676*, 2013.

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

[73] Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166, 2004.

[74] Michael W Graham and David J Miller. Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection. *IEEE Transactions on Signal Processing*, 54(4):1289–1303, 2006.

[75] Yuanhong Li, Ming Dong, and Jing Hua. Simultaneous localized feature selection and model detection for gaussian mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):953–960, 2008.

[76] Masa-Aki Sato. Online model selection based on the variational bayes. *Neural computation*, 13(7):1649–1681, 2001.

[77] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412, 2003.

[78] Ingram Olkin and Thomas A Trikalinos. Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60, 2015.

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

[80] Neil D Lawrence, Christopher M Bishop, and Michael I Jordan. Mixture representations for inference and learning in boltzmann machines. *arXiv preprint arXiv:1301.7393*, 2013.

[81] 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.

[82] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[83] 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.

[84] Wentao Fan and Nizar Bouguila. Online variational learning of generalized dirichlet mixture models with feature selection. *Neurocomputing*, 126:166–179, 2014.

[85] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[86] WHO. accessed September 2018. https://www.who.int/news-room/fact-sheets/detail/cancer.

[87] Muthuraman Alagappan, Jeremy R Glissen Brown, Yuichi Mori, and Tyler M Berzin. Artificial intelligence in gastrointestinal endoscopy: The future is almost here. *World journal of gastrointestinal endoscopy*, 10(10):239, 2018.

[88] Michael T McCann, John A Ozolek, Carlos A Castro, Bahram Parvin, and Jelena Kovacevic. Automated histology analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine*, 32(1):78–87, 2014.

[89] *Colon*, accessed 2009. https://data.broadinstitute.org/bbbc/BBBC018.

[90] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012.

[91] *Zenodo*, accessed 2016. https://zenodo.org/record/53169.

[92] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

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

[94] *ISIC skin dataset*. https://www.isic-archive.com.

[95] *WHO*, accessed Jan 2020. https://www.who.int/malaria/en/.

[96] *NIH*, accessed 2019. https://ceb.nlm.nih.gov/repositories/malaria-datasets/.

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

[98] 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.

[99] Zhanyu Ma and Arne Leijon. Modelling speech line spectral frequencies with dirichlet mixture models. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[100] Itay Mayrose, Nir Friedman, and Tal Pupko. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21(suppl_2):ii151–ii158, 2005.

[101] Yuan Ji, Chunlei Wu, Ping Liu, Jing Wang, and Kevin R Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, 2005.

[102] Anastasios Markitsis and Yinglei Lai. A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646, 2010.

[103] Zhanyu Ma and Andrew E Teschendorff. A variational bayes beta mixture model for feature selection in dna methylation studies. *Journal of bioinformatics and computational biology*, 11(04):1350005, 2013.

[104] Taoufik Bdiri and Nizar Bouguila. Bayesian learning of inverted dirichlet mixtures for svm kernels generation. *Neural Computing and Applications*, 23(5):1443–1458, 2013.

[105] Elise Epaillard and Nizar Bouguila. Data-free metrics for dirichlet and generalized dirichlet mixture-based hmms–a practical study. *Pattern Recognition*, 85:207–219, 2019.

[106] 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.

[107] Elise Epaillard and Nizar Bouguila. Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognition*, 55:125–136, 2016.

[108] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[109] Can Hu, Wentao Fan, Ji-Xiang Du, and Nizar Bouguila. A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, 333:110–123, 2019.

[110] Hirotugu Akaike. Factor analysis and aic. In *Selected papers of hirotugu akaike*, pages 371–386. Springer, 1987.

[111] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

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

[113] 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.

[114] Carl E Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *Advances in neural information processing systems*, pages 881–888, 2002.

[115] Geoffrey J Krishnan, Thriyambakam Ng, SK Ng, T Krishnan, and GJ Mclachlan. The em algorithm. In *Wiley Series in Probability and Statistics: Applied Probability and Statistics, WileyInterscience*. Citeseer, 1997.

[116] Carl Edward Rasmussen. A practical monte carlo implementation of bayesian learning. In *Advances in Neural Information Processing Systems*, pages 598–604, 1996.

[117] Steve R Waterhouse, David MacKay, and Anthony J Robinson. Bayesian methods for mixtures of experts. In *Advances in neural information processing systems*, pages 351–357, 1996.

[118] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[119] 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.

[120] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[121] Carl Edward Rasmussen. A practical monte carlo implementation of bayesian learning. In *Advances in Neural Information Processing Systems*, pages 598–604, 1996.

[122] Hagai Attias. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.

[123] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[124] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[125] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412, 2003.

[126] Ingram Olkin and Thomas A Trikalinos. Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60, 2015.

[127] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[128] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[129] Geoffrey J McLachlan. Mixture models in statistics. 2015.

[130] Wentao Fan and Nizar Bouguila. Variational learning for dirichlet process mixtures of dirichlet distributions and applications. *Multimedia tools and applications*, 70(3):1685–1702, 2014.

[131] Hagai Attias. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.

[132] M Opper and D Saad. Advanced mean field methods: theory and practice. neural information processing series, 2001.

[133] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[134] 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.

[135] 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.

[136] Masa-Aki Sato. Online model selection based on the variational bayes. *Neural computation*, 13(7):1649–1681, 2001.

[137] Wentao Fan and Nizar Bouguila. Online learning of a dirichlet process mixture of generalized dirichlet distributions for simultaneous clustering and localized feature selection. In *Asian Conference on Machine Learning*, pages 113–128, 2012.

[138] Wentao Fan and Nizar Bouguila. Online variational learning for a dirichlet process mixture of dirichlet distributions and its application. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 362–367. IEEE, 2012.

[139] Wentao Fan and Nizar Bouguila. Video background subtraction using online infinite dirichlet mixture models. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013.

[140] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

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

[142] https://www.who.int/news-room/fact-sheets/detail/cancer.

[143] Qiang Li and Robert M Nishikawa. *Computer-aided detection and diagnosis in medical imaging*. Taylor & Francis, 2015.

[144] M Emre Celebi, QUAN Wen, HITOSHI Iyatomi, KOUHEI Shimizu, Huiyu Zhou, and Gerald Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. *Dermoscopy image analysis*, 10:97–129, 2015.

[145] *Skin dataset*. https://challenge.kitware.com.

[146] Nizar Ahmed, Altug Yigit, Zerrin Isik, and Adil Alpkocak. Identification of leukemia subtypes from microscopic images using convolutional neural network. *Diagnostics*, 9(3):104, 2019.

[147] *Leukemia dataset*. https://wiki.cancerimagingarchive.net.

[148] *Bone dataset*. https://wiki.cancerimagingarchive.net.

[149] Reda Chefira and Said Rakrak. A knowledge extraction pipeline between supervised and unsupervised machine learning using gaussian mixture models for anomaly detection. *Journal of Computing Science and Engineering*, 15(1):1–17, 2021.

[150] Yingying Zhu, Youbao Tang, Yuxing Tang, Daniel C Elton, Sungwon Lee, Perry J Pickhardt, and Ronald M Summers. Cross-domain medical image translation by shared latent gaussian mixture model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 379–389. Springer, 2020.

[151] Siva Rajesh Kasa, Sakyajit Bhattacharya, and Vaibhav Rajan. Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping. *Bioinformatics*, 36(2):621–628, 2020.

[152] Yang Zhao, Abhishek K Shrivastava, and Kwok Leung Tsui. Regularized gaussian mixture model for high-dimensional clustering. *IEEE transactions on cybernetics*, 49(10):3677–3688, 2018.

[153] Ian C McDowell, Dinesh Manandhar, Christopher M Vockley, Amy K Schmid, Timothy E Reddy, and Barbara E Engelhardt. Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS computational biology*, 14(1):e1005896, 2018.

[154] Kehua Li, Zhenjun Ma, Duane Robinson, and Jun Ma. Identification of typical building daily electricity usage profiles using gaussian mixture model-based clustering and hierarchical clustering. *Applied energy*, 231:331–342, 2018.

[155] Zexuan Ji, Yong Xia, Quansen Sun, Qiang Chen, and Dagan Feng. Adaptive scale fuzzy local gaussian mixture model for brain mr image segmentation. *Neurocomputing*, 134:60–69, 2014.

[156] Zhanyu Ma, Pravin Kumar Rana, Jalil Taghia, Markus Flierl, and Arne Leijon. Bayesian estimation of dirichlet mixture model with variational inference. *Pattern Recognition*, 47(9):3143–3157, 2014.

[157] Junyang Chen, Zhiguo Gong, and Weiwen Liu. A dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*, 50(5):1609–1619, 2020.

[158] Akram Edalati-rad and Mohammad Mosleh. Improving brain tumor diagnosis using mri segmentation based on collaboration of beta mixture model and learning automata. *Arabian Journal for Science and Engineering*, 44(4):2945–2957, 2019.

[159] Nurvita Trianasari, I Sumertajaya, I Wayan Mangku, et al. Bivariate beta mixture model with correlations. *Commun. Math. Biol. Neurosci.*, 2021:Article–ID, 2021.

[160] S Anuradha and CH Satyanarayana. Medical image segmentation based on beta mixture distribution for effective identification of lesions. In *Recent Developments in Intelligent Computing, Communication and Devices*, pages 133–140. Springer, 2017.

[161] Alberto Llera, Ismael Huertas, Pablo Mir, and Christian F Beckmann. Quantitative intensity harmonization of dopamine transporter spect images using gamma mixture models. *Molecular imaging and biology*, 21(2):339–347, 2019.

[162] Xu Chunyan, Song Yuqing, Liu Zhe, and Bao Xiang. A medical image fusion algorithm based on contourlet transform and t mixture models. *Journal of Nanjing Normal University (Natural Science Edition)*, page 01, 2017.

[163] Zhe Min, Li Liu, and Max Q-H Meng. Generalized non-rigid point set registration with hybrid mixture models considering anisotropic positional uncertainties. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 547–555. Springer, 2019.

[164] Ri-Gui Zhou and Wei Wang. Online nonparametric bayesian analysis of parsimonious gaussian mixture models and scenes clustering. *ETRI Journal*, 43(1):74–81, 2021.

[165] Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature communications*, 8(1):1–11, 2017.

[166] Hui Liu, Zhu Duan, Chao Chen, and Haiping Wu. A novel two-stage deep learning wind speed forecasting method with adaptive multiple error corrections and bivariate dirichlet process mixture model. *Energy Conversion and Management*, 199:111975, 2019.

[167] Halid Ziya Yerebakan and Murat Dundar. Partially collapsed parallel gibbs sampler for dirichlet process mixture models. *Pattern Recognition Letters*, 90:22–27, 2017.

[168] Tadahiro Taniguchi, Ryo Yoshino, and Toshiaki Takano. Multimodal hierarchical dirichlet process-based active perception by a robot. *Frontiers in neurorobotics*, 12:22, 2018.

[169] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):256–270, 2014.

[170] Dingcheng Li, Siamak Zamani, Jingyuan Zhang, and Ping Li. Integration of knowledge graph embedding into topic modeling with hierarchical dirichlet process. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 940–950, 2019.

[171] Takashi Fuse and Keita Kamiya. Statistical anomaly detection in human dynamics monitoring using a hierarchical dirichlet process hidden

markov model. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3083–3092, 2017.

[172] Jia-Ching Wang, Yuan-Shan Lee, Yu-Hao Chin, Ying-Ren Chen, and Wen-Chi Hsieh. Hierarchical dirichlet process mixture model for music emotion recognition. *IEEE Transactions on Affective Computing*, 6(3):261–271, 2015.

[173] Liat Shenhav, Mike Thompson, Tyler A Joseph, Leah Briscoe, Ori Furman, David Bogumil, Itzhak Mizrahi, Itsik Pe'er, and Eran Halperin. Feast: fast expectation-maximization for microbial source tracking. *Nature Methods*, 16(7):627–632, 2019.

[174] Nima Sammaknejad, Yujia Zhao, and Biao Huang. A review of the expectation maximization algorithm in data-driven process identification. *Journal of process control*, 73:123–136, 2019.

[175] Hunter Glanz and Luis Carvalho. An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis*, 167:31–48, 2018.

[176] Ying Shen, Lizhu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. Cbn: Constructing a clinical bayesian network based on data from the electronic medical record. *Journal of biomedical informatics*, 88:1–10, 2018.

[177] Qingping Zhou, Tengchao Yu, Xiaoqun Zhang, and Jinglai Li. Bayesian inference and uncertainty quantification for medical image reconstruction with poisson data. *SIAM Journal on Imaging Sciences*, 13(1):29–52, 2020.

[178] Geoffrey Jones, Neil T Clancy, Yusuf Helo, Simon Arridge, Daniel S Elson, and Danail Stoyanov. Bayesian estimation of intrinsic tissue oxygenation and perfusion from rgb images. *IEEE transactions on medical imaging*, 36(7):1491–1501, 2017.

[179] Alexis Bellot and Mihaela Van Der Schaar. Flexible modelling of longitudinal medical data: A bayesian nonparametric approach. *ACM Transactions on Computing for Healthcare*, 1(1):1–15, 2020.

[180] Guanyang Wang et al. A fast mcmc algorithm for the uniform sampling of binary matrices with fixed margins. *Electronic Journal of Statistics*, 14(1):1690–1706, 2020.

[181] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021.

[182] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[183] Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei. Variational sequential monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 968–977. PMLR, 2018.

[184] Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

[185] Yixin Wang, Andrew C Miller, and David M Blei. Comment: Variational autoencoders as empirical bayes. *Statistical Science*, 34(2):229–233, 2019.

[186] Dustin Tran, Rajesh Ranganath, and David M Blei. Hierarchical implicit models and likelihood-free variational inference. *arXiv preprint arXiv:1702.08896*, 2017.

[187] Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. *International journal of computer assisted radiology and surgery*, 14(3):451–461, 2019.

[188] Ingram Olkin and Thomas A Trikalinos. Constructions for a bivariae beta distribution. *Statistics & Probability Letters*, 96:54–60, 2015.

[189] Zach Dietz, William Lippitt, and Sunder Sethuraman. Stick-breaking processes, clumping, and markov chain occupation laws. *Sankhya A*, pages 1–43, 2021.

[190] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[191] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

[192] 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.

[193] Wentao Fan, Hassen Sallay, Nizar Bouguila, and Sami Bourouis. Variational learning of hierarchical infinite generalized dirichlet mixture models and applications. *Soft Computing*, 20(3):979–990, 2016.

[194] Wentao Fan and Nizar Bouguila. Online data clustering using variational learning of a hierarchical dirichlet process mixture of dirichlet distributions. In *International Conference on Database Systems for Advanced Applications*, pages 18–32. Springer, 2014.

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

[196] Crispian Scully, Jose Bagan, et al. Oral squamous cell carcinoma overview. *Oral oncology*, 45(4/5):301–308, 2009.

[197] Joao Massano, Frederico S Regateiro, Gustavo Januário, and Artur Ferreira. Oral squamous cell carcinoma: review of prognostic and predictive factors. *Oral surgery, oral medicine, oral pathology, oral radiology, and endodontology*, 102(1):67–76, 2006.

[198] Anastasios K Markopoulos. Current aspects on oral squamous cell carcinoma. *The open dentistry journal*, 6:126, 2012.

[199] World Health Organisation. Oral cancer. `https://www.who.int/news-room/fact-sheets/detail/oral-health`, 2020.

[200] Anna G Zygogianni, George Kyrgias, Petros Karakitsos, Amanta Psyrri, John Kouvaris, Nikolaos Kelekis, and Vassilis Kouloulias. Oral

squamous cell cancer: early detection and the role of alcohol and smoking. *Head & neck oncology*, 3(1):2, 2011.

[201] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

[202] Rahman T Y. A histopathological image repository of normal epithelium of oral cavity and oral squamous cell carcinoma. `https://data.mendeley.com/`, 2019.

[203] Patrick P Lin and Shreyaskumar Patel. Osteosarcoma. In *Bone Sarcoma*, pages 75–97. Springer, 2013.

[204] Dela Cruz, C Jennifer, Leonardo C Castor, Celine Margaret T Mendoza, B Arvin Jay, L Song Cherry Jane, P Torres Bailey Brian, et al. Determination of blood components (wbcs, rbcs, and platelets) count in microscopic images using image processing and analysis. In *2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pages 1–7. IEEE, 2017.

[205] https://www.kaggle.com/paultimothymooney/blood-cells.

[206] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138, 2021.

[207] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

[208] Gunasekaran Manogaran, V Vijayakumar, R Varatharajan, Priyan Malarvizhi Kumar, Revathi Sundarasekar, and Ching-Hsien Hsu. Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering. *Wireless personal communications*, 102(3):2099–2116, 2018.

[209] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[210] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. *arXiv preprint arXiv:1301.6676*, 2013.

[211] David M Blei and Michael I Jordan. Variational methods for the dirichlet process. In *Proceedings of the twenty-first international conference on Machine learning*, page 12. ACM, 2004.

[212] Christopher M Bishop. Variational learning in graphical models and neural networks. In *International Conference on Artificial Neural Networks*, pages 13–22. Springer, 1998.

[213] Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

[214] Zhanyu Ma and Arne Leijon. Expectation propagation for estimating the parameters of the beta distribution. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2082–2085. IEEE, 2010.

[215] Jaak Panksepp. Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist. *Perspectives on psychological science*, 2(3):281–296, 2007.

[216] Iris B Mauss and Michael D Robinson. Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237, 2009.

[217] Paolo Fornacciari, Stefano Cagnoni, Monica Mordonini, Leonardo Tarollo, and Michele Tomaiuolo. Application of lovheim model for emotion detection in english tweets. In *WOA*, pages 149–155, 2019.

[218] Hao Chao, Liang Dong, Yongli Liu, and Baoyun Lu. Emotion recognition from multiband eeg signals using capsnet. *Sensors*, 19(9):2212, 2019.

[219] Xiaofen Xing, Zhenqi Li, Tianyuan Xu, Lin Shu, Bin Hu, and Xiangmin Xu. Sae+ lstm: A new framework for emotion recognition from multichannel eeg. *Frontiers in neurorobotics*, 13:37, 2019.

[220] Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience*, 2020, 2020.

[221] https://www.kaggle.com/birdy654/eeg-brainwave-dataset-mental-state.

[222] Jordan J Bird, Luis J Manso, Eduardo P Ribeiro, Aniko Ekart, and Diego R Faria. A study on mental state classification using eeg-based brain-machine interface. In *2018 International Conference on Intelligent Systems (IS)*, pages 795–800. IEEE, 2018.

[223] Jordan J Bird, A Ekart, CD Buckingham, and Diego R Faria. Mental emotional sentiment classification with an eeg-based brain-machine interface. In *Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP'19)*, 2019.

[224] https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones#.

[225] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. *Esann*, 3:3, 2013.

[226] Matthias Boeker, Michael A Riegler, Hugo L Hammer, Pål Halvorsen, Ole Bernt Fasmer, and Petter Jakobsen. Diagnosing schizophrenia from activity records using hidden markov model parameters. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 432–437. IEEE, 2021.

[227] Nelson F Monroy and Miguel Altuve. Hidden markov model-based heartbeat detector using different transformations of ecg and abp signals. In *15th International Symposium on Medical Information Processing and Analysis*, volume 11330, page 113300S. International Society for Optics and Photonics, 2020.

[228] Qi Huang, Dwayne Cohen, Sandra Komarzynski, Xiao-Mei Li, Pasquale Innominato, Francis Lévi, and Bärbel Finkenstädt. Hidden

markov models for monitoring circadian rhythmicity in telemetric activity data. *Journal of The Royal Society Interface*, 15(139):20170885, 2018.

[229] Nelson F Monroy and Miguel Altuve. Joint exploitation of hemodynamic and electrocardiographic signals by hidden markov models for heartbeat detection. In *Latin American Conference on Biomedical Engineering*, pages 208–217. Springer, 2019.

[230] Jaeah Kim, Shashank Singh, Erik D Thiessen, and Anna V Fisher. A hidden markov model for analyzing eye-tracking of moving objects. *Behavior research methods*, 52(3):1225–1243, 2020.

[231] Min Wang, Sherif Abdelfattah, Nour Moustafa, and Jiankun Hu. Deep gaussian mixture-hidden markov model for classification of eeg signals. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(4):278–287, 2018.

[232] Amrit Dhar, Duncan K Ralph, Vladimir N Minin, and Frederick A Matsen IV. A bayesian phylogenetic hidden markov model for b cell receptor sequence analysis. *PLoS computational biology*, 16(8):e1008030, 2020.

[233] Hojat Ghimatgar, Kamran Kazemi, Mohammad Sadegh Helfroush, and Ardalan Aarabi. An automatic single-channel eeg-based sleep stage scoring method based on hidden markov model. *Journal of neuroscience methods*, 324:108320, 2019.

[234] Gemeng Zhang, Biao Cai, Aiying Zhang, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu-Ping Wang. Estimating dynamic functional brain connectivity with a sparse hidden markov model. *IEEE transactions on medical imaging*, 39(2):488–498, 2019.

[235] Gunasekaran Manogaran, V Vijayakumar, R Varatharajan, Priyan Malarvizhi Kumar, Revathi Sundarasekar, and Ching-Hsien Hsu. Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering. *Wireless personal communications*, 102(3):2099–2116, 2018.

[236] Rozita Rastghalam, Habibollah Danyali, Mohammad Sadegh Helfroush, M Emre Celebi, and Mojgan Mokhtari. Skin melanoma detection in microscopic images using hmm-based asymmetric analysis and expectation maximization. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3486–3497, 2021.

[237] Shruti Sharma and Munish Rattan. An improved segmentation and classifier approach based on hmm for brain cancer detection. *The Open Biomedical Engineering Journal*, 13(1), 2019.

[238] Cecile JA Wolfs, Nicolas Varfalvy, Richard AM Canters, Sebastiaan MJJG Nijsten, Djoya Hattu, Louis Archambault, and Frank Verhaegen. External validation of a hidden markov model for gamma-based classification of anatomical changes in lung cancer patients using epid dosimetry. *Medical Physics*, 47(10):4675–4682, 2020.

[239] Mohammadreza Momenzadeh, Mohammadreza Sehhati, and Hossein Rabbani. Using hidden markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles. *Journal of Biomedical Informatics*, 111:103570, 2020.

[240] Hong Zheng, Ruoyin Wang, Wencheng Xu, Yifan Wang, and Wen Zhu. Combining a hmm with a genetic algorithm for the fault diagnosis of photovoltaic inverters. *Journal of power electronics*, 17(4):1014–1026, 2017.

[241] Hongfa Ding, Youliang Tian, Changgen Peng, Youshan Zhang, and Shuwen Xiang. Inference attacks on genomic privacy with an improved hmm and an rcnn model for unrelated individuals. *Information Sciences*, 512:207–218, 2020.

[242] Hassan Satori, Ouissam Zealouk, Khalid Satori, and Fatima El-Haoussi. Voice comparison between smokers and non-smokers using hmm speech recognition system. *International Journal of Speech Technology*, 20(4):771–777, 2017.

[243] Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition. *Speech Communication*, 108:15–32, 2019.

[244] José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, and Néstor Becerra Yoma. Dnn-hmm based automatic speech recognition for hri scenarios. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 150–159, 2018.

[245] Thomas Schatz and Naomi H Feldman. Neural network vs. hmm speech recognition systems as models of human cross-linguistic phonetic perception. In *Proceedings of the conference on cognitive computational neuroscience*, 2018.

[246] José Novoa, Josué Fredes, Víctor Poblete, and Néstor Becerra Yoma. Uncertainty weighting and propagation in dnn–hmm-based speech recognition. *Computer Speech & Language*, 47:30–46, 2018.

[247] Rabeet Fatmi, Sherif Rashad, and Ryan Integlia. Comparing ann, svm, and hmm based machine learning methods for american sign language recognition using wearable motion sensors. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0290–0297. IEEE, 2019.

[248] Akram Emdadi and Changiz Eslahchi. Auto-hmm-lmf: feature selection based method for prediction of drug response via autoencoder and hidden markov model. *BMC bioinformatics*, 22(1):1–22, 2021.

[249] Jian-Hua Zhang, Xiu-Ling Liu, Zheng-Li Hu, Yi-Lun Ying, and Yi-Tao Long. Intelligent identification of multi-level nanopore signatures for accurate detection of cancer biomarkers. *Chemical Communications*, 53(73):10176–10179, 2017.

[250] Agastya Silvina, Juliana Bowles, and Peter Hall. On predicting the outcomes of chemotherapy treatments in breast cancer. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 180–190. Springer, 2019.

[251] Md Zia Uddin. Human activity recognition using segmented body part and body joint features with hidden markov models. *Multimedia Tools and Applications*, 76(11):13585–13614, 2017.

[252] Zhelong Wang and Ye Chen. Recognizing human concurrent activities using wearable sensors: a statistical modeling approach based on parallel hmm. *Sensor Review*, 2017.

[253] Mariana Abreu, Marília Barandas, Ricardo Leonardo, and Hugo Gamboa. Detailed human activity recognition based on multiple hmm. In *BIOSIGNALS*, pages 171–178, 2019.

[254] Guangxin Liu, Yuru Kang, and Huichao Men. Char-hmm: An improved continuous human activity recognition algorithm based on hidden markov model. In *Mobile Ad-hoc and Sensor Networks: 13th International Conference*, volume 747, pages 271–282. Springer, 2018.

[255] Xin Tong, Yan Su, Zhaofeng Li, Chaowei Si, Guowei Han, Jin Ning, and Fuhua Yang. A double-step unscented kalman filter and hmm-based zero-velocity update for pedestrian dead reckoning using mems sensors. *IEEE Transactions on Industrial Electronics*, 67(1):581–591, 2019.

[256] Georgia Chalvatzaki, Xanthi S Papageorgiou, Costas S Tzafestas, and Petros Maragos. Estimating double support in pathological gaits using an hmm-based analyzer for an intelligent robotic walker. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 101–106. IEEE, 2017.

[257] Shuo Yu, Hsinchun Chen, and Randall A Brown. Hidden markov model-based fall detection with motion sensor orientation calibration: A case for real-life home monitoring. *IEEE journal of biomedical and health informatics*, 22(6):1847–1853, 2017.

[258] Xiang Chen, Zhi-Xin Wang, and Xian-Ming Pan. Hiv-1 tropism prediction by the xgboost and hmm methods. *Scientific reports*, 9(1):1–8, 2019.

[259] Shing-Tai Pan and Wei-Ching Li. Fuzzy-hmm modeling for emotion detection using electrocardiogram signals. *Asian Journal of Control*, 22(6):2206–2216, 2020.

[260] Xingce Wang, Yue Liu, Zhongke Wu, Xiao Mou, Mingquan Zhou, Miguel A González Ballester, and Chong Zhang. Automatic labeling of

vascular structures with topological constraints via hmm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 208–215. Springer, 2017.

[261] Shadi AlZu'bi, Sokyna AlQatawneh, Mohammad ElBes, and Mohammad Alsmirat. Transferable hmm probability matrices in multi-orientation geometric medical volumes segmentation. *Concurrency and Computation: Practice and Experience*, 32(21):e5214, 2020.

[262] SN Kumar, S Muthukumar, H Kumar, P Varghese, et al. A voyage on medical image segmentation algorithms. *Biomedical Research (0970-938X)*, 2018.

[263] Cheol-Hong Min. Automatic detection and labeling of self-stimulatory behavioral patterns in children with autism spectrum disorder. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 279–282. IEEE, 2017.

[264] Preetam Srikar Dammu and Raju Surampudi Bapi. Temporal dynamics of the brain using variational bayes hidden markov models: Application in autism. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 121–130. Springer, 2019.

[265] Meenakshi Chatterjee, Nikolay V Manyakov, Abigail Bangerter, Dzmitry A Kaliukhovich, Shyla Jagannatha, Seth Ness, and Gahan Pandina. Learning scan paths of eye movement in autism spectrum disorder. In *Digital Personalized Health and Medicine*, pages 287–291. IOS Press, 2020.

[266] S Priyadharshini and K Sivaranjani. Investigating and statistical analysis of autism spectrum disorders: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(7):13–15, 2017.

[267] Jeroen Van Schependom, Diego Vidaurre, Lars Costers, Martin Sjøgård, Marie B D'hooghe, Miguel D'haeseleer, Vincent Wens, Xavier De Tiège, Serge Goldman, Mark Woolrich, et al. Altered transient brain dynamics in multiple sclerosis: Treatment or pathology? *Human brain mapping*, 40(16):4789–4800, 2019.

[268] Nazanin Esmaili, Massimo Piccardi, Bernie Kruger, and Federico Girosi. Analysis of healthcare service utilization after transport-related injuries by a mixture of hidden markov models. *PLoS one*, 13(11):e0206274, 2018.

[269] BalaAnand Muthu, CB Sivaparthipan, Gunasekaran Manogaran, Revathi Sundarasekar, Seifedine Kadry, A Shanthini, and Antony Dasel. Iot based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-peer networking and applications*, 13(6):2123–2134, 2020.

[270] Alexandre Vimont, Henri Leleu, and Isabelle Durand-Zaleski. Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in france. *The European Journal of Health Economics*, pages 1–13, 2021.

[271] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[272] E Fox, E Sudderth, M Jordan, and A Willsky. Developing a tempered hdp-hmm for systems with state persistence. *MIT LIDS*, 2007.

[273] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An hdp-hmm for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319, 2008.

[274] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

[275] Ava Bargi, Richard Yi Da Xu, and Massimo Piccardi. An online hdp-hmm for joint action segmentation and classification in motion capture data. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7. IEEE, 2012.

[276] Natraj Raman and Stephen J Maybank. Action classification using a discriminative multilevel hdp-hmm. *Neurocomputing*, 154:149–161, 2015.

[277] Ava Bargi, Richard Yi Da Xu, and Massimo Piccardi. Adon hdp-hmm: an adaptive online model for segmentation and classification of sequential data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):3953–3968, 2017.

[278] Jing Zhao, Samanvitha Basole, and Mark Stamp. Malware classification with gmm-hmm models. *arXiv preprint arXiv:2103.02753*, 2021.

[279] Fengquan Zhang, Songyang Han, Huaming Gao, and Taipeng Wang. A gaussian mixture based hidden markov model for motion recognition with 3d vision device. *Computers & Electrical Engineering*, 83:106603, 2020.

[280] Yufei Li, Bo Hu, Tao Niu, Shengpu Gao, Jiahao Yan, Kaigui Xie, and Zhouyang Ren. Gmm-hmm-based medium-and long-term multi-wind farm correlated power output time series generation method. *IEEE Access*, 9:90255–90267, 2021.

[281] Xiaoyan Cheng, Binke Huang, and Jing Zong. Device-free human activity recognition based on gmm-hmm using channel state information. *IEEE Access*, 2021.

[282] Ching Leng Peter Lim, Wai Lok Woo, Satnam S Dlay, and Bin Gao. Heartrate-dependent heartwave biometric identification with thresholding-based gmm–hmm methodology. *IEEE Transactions on Industrial Informatics*, 15(1):45–53, 2018.

[283] Le Chen, David Barber, and Jean-Marc Odobez. Dynamical dirichlet mixture model. Technical report, IDIAP, 2007.

[284] Rim Nasfi, Manar Amayri, and Nizar Bouguila. A novel approach for modeling positive vectors with inverted dirichlet-based hidden markov models. *Knowledge-Based Systems*, 192:105335, 2020.

[285] Elise Epaillard and Nizar Bouguila. Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection. *IEEE transactions on neural networks and learning systems*, 30(4):1034–1047, 2018.

[286] Aonan Zhang, San Gultekin, and John Paisley. Stochastic variational inference for the hdp-hmm. In *Artificial Intelligence and Statistics*, pages 800–808. PMLR, 2016.

[287] Yixin Wang and David Blei. Variational bayes under model misspecification. *Advances in Neural Information Processing Systems*, 32:13357–13367, 2019.

[288] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.*, 32(24):18069–18083, 2020.

[289] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230, 1973.

[290] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[291] Tzu-Tsung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

[292] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[293] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[294] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. learning in graphical models. *MI Jordan (ed)*, 105162, 1999.

[295] 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.

[296] John Paisley, Student Member, and Lawrence Carin. Hidden markov models with stick breaking priors, to appear. *IEEE Trans. on Sig. Proc*, 2009.

[297] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[298] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE, 2010.

[299] Hesam Sagha, Sundara Tejaswi Digumarti, José del R Millán, Ricardo Chavarriaga, Alberto Calatroni, Daniel Roggen, and Gerhard Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 36–40. IEEE, 2011.