A Study On Online Variational Learning : Medical Applications

Meeta Kalra

A Thesis
In
The Concordia Institute for Information Systems Engineering

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

January 2020

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Meeta Kalra**

Entitled: **A Study On Online Variational Learning : Medical Applications**

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

Master of Applied Science (Information Systems Security)

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

Signed by the final examining commitee:

Dr. Chun Wang _____ Chair

Dr. Nizar Bouguila _____ Supervisor

Dr. Mohsen Ghafouri _____ CIISE Examiner

Dr. Hoi Dick Ng _____ External Examiner

Approved _____

Dr. Mohammad Mannan, Graduate Program Director

2019.12.17 _____

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

Abstract

A Study On Online Variational Learning : Medical Applications

Meeta Kalra

Data mining is an extensive area of research which is applied in various critical domains. In clinical aspect, data mining has emerged to assist clinicians in early detection, diagnosis and prevention of diseases. On the other hand, advances in computational methods have led to the implementation of machine learning in multi-modal clinical image analysis such as in CT, X-ray, MRI, microscopy among others. A challenge to these applications is the high variability, inconsistent regions with missing edges, absence of texture contrast and high noise in the background of biomedical images. To overcome this limitation various segmentation approaches have been investigated to address these shortcomings and to transform medical images into meaningful information. It is of utmost importance to have the right match between the bio-medical data and the applied algorithm.

During the past decade, finite mixture models have been revealed to be one of the most flexible and popular approaches in data clustering. Here, we propose a statistical framework for online variational learning of finite mixture models for clustering medical images. The online variational learning framework is used to estimate the parameters and the number of mixture components simultaneously in a unified framework, thus decreasing the computational complexity of the model and the over fitting problems experienced with maximum likelihood approaches guaranteeing convergence. In online learning, the data becomes available in a sequential order, thus sequentially updating the best predictor for the future data at each step, as opposed to batch learning techniques which generate the best predictor by learning the entire data set at once. The choice of distributions remains the core concern of mixture models in recent research. The efficiency of Dirichlet family of distributions for this purpose has been proved in latest studies especially for non-Gaussian data. This led us to analyze online variational learning approach for finite mixture models based on different distributions.

To this end, our contribution is the application of online variational learning approach to design finite mixture models based on inverted Dirichlet, generalized inverted Dirichlet with feature selection and inverted Beta-Liouville distributions in medical domain. We evaluated our proposed models on different biomedical image data sets. Furthermore, in each case we compared the proposed algorithm with other popular algorithms. The models detect the disease patterns with high confidence. Computational and statistical approaches like the ones presented in our work hold a significant impact on medical image analysis and interpretation in both clinical applications and scientific research. We believe that the proposed models have the capacity to address multi modal biomedical image data sets and can be further applied by researchers to analyse correct disease patterns.

Acknowledgments

Two and half years ago, I moved to Canada and took a deep dive into my passion for data science by pursuing my Masters degree. As quoted by Rabindranath Tagore ”Everything comes to us that belongs to us if we create the capacity to receive it.” This is how I would like to express my profound gratitude to my supervisor **Prof Dr. Nizar Bouguila**. I started my journey with Concordia University in a different department when I took my mentor's course as an elective. I was so impressed by his extraordinary teaching qualities and knowledge in data mining that I approached him to do a research based master under him. I would be forever grateful to him for giving me by far the best opportunity to work under his guidance to pursue my passion. I would use this opportunity to express my deepest appreciation for his persistent guidance and endless support, patience, motivation and encouragement during the tough phases of my journey here. I cannot thank him enough for being an outstanding and extraordinary supervisor who helped me achieve my dreams.

Besides my advisor, a special thanks to Dr. Michael Osadebey for sharing his knowledge in the field of medical sciences and showing me the right path in order to constantly progress in my research.

A heartiest thanks to Mr. Muhammad Azam who is like a big brother to me here. He always made sure that I was in high spirits and kept me motivated whenever I felt low.

I am also grateful to Kamal, Hieu and Jaspreet who were the always there to help me in the times of struggle and have coffee together to de-stress. Hieu, thank you for checking on me every week with ”How's Progress” which helped me motivate myself each day.

I have been really lucky to have such awesome lab mates who are like a family now since they made my two years journey unforgettable. I would like to thank Narges, Nuha, Maryam, Samr, Omar, Basim, Omayama, Pantea, Sunny, Ornela and other lab members, who have always shared their vast knowledge as well as taken the time and effort to help me keep going with the hard work.

It is noteworthy to mention my gratitude and love to the special Concordia University Faculty, Jennifer Drummond, Lindsay Faul, Jewel Perlin, Howard Magonet, Jasia Stuart and Tailor to help me through the tough times I confronted. I wouldn't have come this far without having either of you in this journey.

A special token of thanks to Lyne Denis and Ariana Hipsagh who treated me like their daughter and were always there to provide affection, love, and guidance.

I would also like extend my immense love and gratitude to my roommates, Shuting and Chaio for always helping me stay positive when I was away from family.

I would also like to express my love and thank you to all my friends far and near especially my best friends from India, Sejal, Akrati, Apurva and Nazim for sticking with me through all the stages of my ups and downs and pushing me whenever I decided to give up on my dreams over the phone by sending tons of love through calls, texts and gifs.

Lastly, immense gratitude to my parents Dinesh and Vanita for being the absolute chill parents I could get and for sending me to Montreal to accomplish my dreams. Your catching up to technology for your daughters and amazing whatsapp calls and emoticons never made me feel less loved. I would not be here today if it were not for you, for your vision and the sacrifices you have made for us. Thank you for the immense love, support, guidance and continuous encouragement. Dr. Priyata, thank you for being the best elder sister I could have ever wanted. There is no better friend than having you as a sister who has been an idol, inspiration and strength for me since the time I came into this world. Thanks for being my constant teacher for not just to teach me about life but also to go through each of my essays which I wrote in childhood to the publications I wrote for my Master's till date. You all were and have been my pillars of faith, hope, courage and strength. If there is anything that stayed constant in 27 years of my life it was you three who are literally -my world.

Contents

List of Figures **ix**

List of Tables **xii**

1 Introduction **1**

1.1 Background 1

1.2 Data mining and its use in healthcare 4

1.2.1 Classification 4

1.2.2 Trend Analysis 5

1.2.3 Clustering 5

1.2.4 Regression 6

1.2.5 Association 6

1.2.6 Summarization 6

1.3 Contributions 6

1.4 Thesis Overview 8

2 Online Variational learning for Finite Inverted Dirichlet Mixture Model **9**

2.1 Model Specification 9

2.1.1 Finite Inverted Dirichlet Mixture Model 9

2.2 Online variational learning for finite inverted Dirichlet mixture model . . . 12

2.2.1 Variational inference 12

2.2.2 Online variational inference 15

2.3 Experimental results 19

2.3.1 Synthetic data 20

2.3.2 Medical image data sets 22

3 Online Variational learning using Finite Generalized Inverted Dirichlet Mixture Model with Feature Selection 30
3.1 Model specification . . . 30
3.1.1 Prior Specifications . . . 33
3.2 Online variational learning for finite generalized inverted Dirichlet mixture mode with feature selection . . . 34
3.3 Experimental results . . . 43
3.3.1 Image segmentation . . . 44
3.3.2 Synthetic data . . . 45
3.3.3 Medical image data sets . . . 46

4 Online Variational learning for Finite Inverted Beta-Liouville Mixture Model 55
4.1 Model Specification . . . 55
4.1.1 Finite Inverted Beta-Liouville Mixture Model . . . 55
4.2 Online variational learning for finite Inverted Beta-Liouville Mixture Model 59
4.3 Experimental Results . . . 66
4.3.1 Brain Tumor Detection . . . 67
4.3.2 Diabetic retinopathy (DR) Optic Disc Localization and Detection . 69
4.3.3 Skin Melanoma Detection . . . 71
4.3.4 Colon Cancer Detection . . . 73
4.3.5 Computer Aided Detection (CAD) of Malaria . . . 75

5 Conclusion 77

Bibliography 80
A Appendix . . . 91
A.1 Proof of equation (2.17): Variational solution of $Q(\mathcal{Z})$ . . . 91
A.2 Proof of equation (2.18), (2.22) and (2.23) . . . 92
A.3 Proof of equation (2.27) . . . 94

List of Figures

2.1 Graphical model representation for finite inverted Dirichlet mixture. Symbols in the circle denote the random variables; otherwise, they denote the model parameters. 11
2.2 Results using Jaccard and dice evaluation metrics for brain tumour detection 24
2.3 Mean and standard deviation results for brain tumour detection 24
2.4 Best segmented brain MRI images : A. Input image, B. 3rd Cluster, C. 6th Cluster, D. 7th Cluster 25
2.5 Segmented brain MRI images after post processing: A.Clustered image, B. Binary image, C. Clustered after filling holes, D. Processed clustered image and E. Ground truth image. The data set was taken from BRATS database [1, 2] where the ground truth data was available. 25
2.6 Results using Jaccard and Dice evaluation metrics for skin lesion diagnosis 26
2.7 Mean and standard deviation results for skin lesion diagnosis 26
2.8 Best Segmented Skin Lesion Images: A. Input image, B. 0th Cluster, C. 9th Cluster, D. 14th Cluster E. 10th Cluster 27
2.9 Segmented Skin lesion images after post processing: A.Clustered image, B. Greyscale image, C. Binary image, D. Binary image after filling holes, and E. Ground truth image. The data set was taken from ICIS database where the ground truth data was available. 27
2.10 Results using Jaccard and Dice evaluation metrics for lung tuberculosis detection 28
2.11 Mean and standard deviation results for lung tuberculosis detection 28
2.12 Best segmented Lung images : A. Input image, B. 10th Cluster, C. 7th Cluster, D. 4th Cluster,E. 0th Cluster 29

2.13 Lung X-ray after post processing: A.Clustered image, B. Binary image, C. Clustered after filling holes, D. Processed cluster and E. Ground truth image. The data set was taken from Montgomery County - Chest X-ray Database provided by national library of medicine where the ground truth data was available. 29
3.1 Graphical representation of finite GID mixture model with feature selection. The circles represent the random variables and model parameters. Numbers in the upper right corners of the plates indicate the number of repetitions. 34
3.2 Example of best segmented brain MRI images for patient 1 : A. Input MRI image, B. 7^{th} Cluster Image, C. 8^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image 49
3.3 Example of best segmented brain MRI images for patient 2 : A. Input MRI image, B. 0^{th} Cluster Image, C. 5^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image. 49
3.4 Example of best segmented dermoscopic images for patient 1 : A. Input image, B. 5^{th} Cluster Image, C. 8^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image 51
3.5 Example of best segmented dermoscopic images for patient 2 : A. Input image, B. 0^{th} Cluster Image, C. 5^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image. 51
3.6 Examples of malaria cells labelled as parasitized in the data set 53
3.7 Examples of malaria cells labelled as uninfected in the data set 54
4.1 Four examples of inverted beta liouville distributions 56
4.2 Examples of finite IBL Mixture model with different components 57
4.3 Graphical model representation for finite IBL mixture model. Symbols in the circle denote the random variables; otherwise, they denote the model parameters. 59
4.4 Example of best segmented brain MRI images for LGG. In the full panel from left to right are : A. Input MRI image, B. 2^{nd} Cluster Image, C. 7^{th} Cluster Image, D. Predicted Image after post processing, E. Ground Truth Image. The dice coefficient for this example was 92% The images show FLAIR modality and predict the FLAIR abnormality. 68

4.5 Example of best segmented optic disc (OD) images for OD detection in retinal fundus image. In the full panel, from left to right are : A. Input DR image, B. 10^{th} Cluster Image, C. 15^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image 70
4.6 Example of best segment generated from the implementation OVIBLMM algorithm on dermoscopic images of melanoma. In the full panel, from left to right are : A. Input melanoma image from the data set, B. 12^{th} Cluster Image, C. 14^{th} Cluster Image, D. Predicted Image, E. Ground Truth Image . 72
4.7 Representative best segment of human HT-29 colon cancer cells where the cytoplasm is segmented. In the full panel, from left to right are: A. Input Colon actin image, B. 1^{st} Cluster Image, C. 3^{rd} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image 74

List of Tables

2.1 Real and estimated parameters of different data sets. N denotes the total number of data points, N_j denotes the number of data points in the cluster j. $\alpha_{j1}, \alpha_{j2}, \alpha_{j3}$ and π_j are the real parameters and $\hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \hat{\alpha}_{j3}$ and $\hat{\pi}_j$ are the parameters estimated by our proposed algorithm. 22

3.1 Real and estimated parameters of different data sets. N denotes the total number of data points, N_j denotes the number of data points in the cluster j. $\alpha_{j1}, \beta_{j1}, \alpha_{j2}, \beta_{j2}$ and π_j are the real parameters and $\hat{\alpha}_{j1}, \hat{\beta}_{j1}, \hat{\alpha}_{j2}, \hat{\beta}_{j2},$ and $\hat{\pi}_j$ are the parameters estimated by our proposed algorithm. 46

3.2 Evaluation metrics for brain tumor detection 49

3.3 Evaluation metrics for skin melanoma detection 51

3.4 Evaluation metrics for malaria data set 54

4.1 Evaluation metrics for brain tumor detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were ARI, AMI, V-measure, Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy. 68

4.2 Evaluation metrics for optic disc detection in DR. OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy. 71

4.3 Evaluation metrics for skin melanoma detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy. 72

4.4 Evaluation metrics for human colon cancer detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy. 75
4.5 Evaluation metrics for malaria data set where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Accuracy, Precision, Recall and F1- Score. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy. 76
5.6 Overview of the biomedical data sets on which our models have been implemented in this study. The data set is heterogeneous in nature (various organs and modalities). 78

Chapter 1

Introduction

1.1 Background

In modern era, imaging is increasingly implemented in medical diagnosis and scientific research. Thus, leading to advances in technical and diagnostic improvements in the field of medical imaging [3]. This has resulted in the emergence of medical data mining which is an increasingly notable research domain [4]. In medicine, imaging is a non invasive biomedical technique applied in clinical context to identify and diagnose diseases [5, 6]. Depending on the medical imaging technique used, medical imaging can give insights into two categories of biomedical analysis, structural or functional [7]. For example, MRI can be used to give structural information of the tumour mass but can also be used to monitor blood flow into the tumour, thus giving functional insights [8]. Although, most medical images represent anatomical structure of the body, application of data mining on them can give valuable insights on the physiology and diagnosis for computer aided-diagnosis [9, 10, 11]. However, extraction and analysis of pertinent information from the often noisy medical images is becoming a more and more pressing issue [6, 12]. In this aspect, medical image segmentation applications have great potential for exploiting hidden patterns in these data sets. The purpose of medical image segmentation is to divide an image into several different regions according to the characteristics of regions within that image [13]. However, identifying the pixels for patho-physiological features from the medical images is a challenging task due to high background in such images. Therefore, the implementation of different algorithms which help detecting segments that can assist doctors to diagnose the disease have recently gained tremendous traction in this field. Many researchers have

proposed various automated techniques to address this, for e.g., traditionally edge detection methods along with mathematical modelling was explored. Later on, a combination of machine learning and feature extraction techniques are explored. In this aspect, many algorithms have been applied, such as classification, clustering, association rules, decision trees, and artificial neural networks [14].

Among the many existing clustering methods, finite mixture models have gained increasing interest and have been successfully applied in various fields such as data mining, machine learning, image processing and bioinformatics. This popularity lies on their solutions to model heterogeneous data. In the case of medical images, the data set can be divided into two or more sub-populations (components). These components are defined by the parameters of the mixture that would provide adequate adaptation to the data [15, 16]. From the different kinds of mixture models, Gaussian mixture model has been a popular choice in various studies because of its simplicity [17]. GMM has disseminated the industry and has profound applications in a number of data analysis tasks [18]. However, Gaussian mixture is not the best choice in all applications and fails to discover the actual underlying data structure when partitions are not Gaussian. In this context, the Dirichlet family performs better with proportional data [15, 19]. Recent work has demonstrated through many real-world applications which involve positive vectors that other mixture models could be better alternatives to clustering and data modelling, such as inverted dirichlet (ID), generalized inverted dirichlet (GID), or inverted Beta-Liouville (IBL) mixtures. In particular, data extracted from texts, images, or videos [20, 21]. Several works have been proposed to model positive vectors based on inverted Dirichlet mixture models (IDMM) [19, 22, 23].

The traditional approaches to learn finite mixture models are based on maximum likelihood (ML) [24] which is usually carried out via expectation maximization (EM) [25]. The differences between Monte Carlo Markov Chain (MCMC), online variational learning and maximum likelihood estimation (MLE) methods have been succinctly described [26]. For instance, online variational learning and MLE methods have been described to be more efficient than MCMC. Out of these methods, maximum likelihood estimation (MLE) has been the most well described and well-known in probabilistic models. It has been extensively applied for estimation of parameters in modern statistics. One problem which EM

faces is the over fitting and being unable to determine the model complexity [23]. However, the disadvantage can be offset by the adoption of Bayesian framework. The Bayesian approach is very comprehensive since the posterior distribution covers the uncertainty of the process. In essence, the Bayesian framework goes hand-in-hand with an approximation scheme. Robert and Cabella [27] describe the utilization of MCMC techniques as the most significant sampling methods which enabled the application of Bayesian techniques in wide aspects of studies. However, the critical challenge of MCMC is limitation to small scale applications due to the need of high computational resources to solve it. In addition, convergence diagnosis is very complex to assess. Thus, variational inference method was developed to overcome the limitations of MCMC.

Variational inference, also known as variational Bayes, is a deterministic approximation method, where, the model's posterior distribution is approximated using analytical procedures [28]. It has generated a lot of interest in finite mixture models through the provision of high generalization schemes and high computation tractability. Model selection and parameter estimation can be performed simultaneously through the use of variational inference. Online mixture learning algorithms have been described to be more efficient in the modeling of data streams, as compared to batch algorithms. Examples include online Gaussian Mixture Models (GMM) considered for instance in [29]. Most of the recent past research works have demonstrated that simulations with other methodological approaches can be better than the GMM when dealing with non-gaussian data. For example, Bouguila and Ziou [20] developed an online learning approach in which the MML criterion was utilized and incorporated. An online variational inference algorithm has been developed in [30], also. In this study, we have proposed a more elaborated approach in which model selection and online learning are examined simultaneously.

The developed approach will be explained in detail in Chapter 2. The authors in [31, 32] have developed an efficient application of this model to finite Dirichlet mixture model. Our intention is to study the efficiency of the above mentioned model when applied to inverted Dirichlet (ID), generalized inverted Dirichlet (GID) and inverted Beta-Liouville (IBL) mixture models. We evaluated our proposed models on different biomedical image data sets including optic disc detection and localization in diabetic retinopathy, digital imaging in

melanoma lesion detection and segmentation, brain tumour detection, colon cancer detection and computer aid detection (CAD) of Malaria using different evaluation metrics in each experimental case.

1.2 Data mining and its use in healthcare

In simple terms data mining approach uses computational models to extract useful information from the data. Particularly, in healthcare, the data generated is rich and multi-modal, for e.g., electronic medical records data, medical image data, proteomic and genomic data to name a few [33]. Despite the abundance of data, computer aided decision support is at its nascency. In this aspect, data mining is implemented to assist clinicians in the early detection, diagnosis and prevention of diseases. This is achieved by establishing models on medical data sets. These models learn from the data and help predict disease prognosis and progression. Basically, data mining models are grouped into two categories; descriptive and predictive models [34]. As the name suggests, descriptive models define the associations that are represented in the data by pattern discovery [35]. In contrast, predictive models are applied to predict a future behaviour or trend as opposed to giving information of the existing behaviour [36]. Depending on the type of medical data, a descriptive or predictive model is chosen for. The important data mining tasks applied in healthcare is to validate conclusions on the diagnosis and treatment regimes [37, 38] are:

1.2.1 Classification

Classification techniques are largely based on statistical models. As per the name, classification refers to the concept of assigning data into target classes. Data are grouped into testing and training sets whenever classification is being implemented. Training data are used by the classifiers in coming up with conclusive attributes of the data before they are put in classes whereas the testing data sets are used to determine the correctness or accuracy of the classifier.

In hospitals or clinics, classification can be applied to determine risk pattern of each patient depending on the data that are stored about the patient [39]. Since these classifiers are rule based, they are implemented to classify the patient into low or high risk populations for a certain diagnosis or disease [40]. In this approach, the patient cases are known thus,

classification can be described as supervised learning. A practical application of classification is that the hospitals and diagnosing units determine the cost of treating the patients in the classes of low risk or high risk diseases [41].

1.2.2 Trend Analysis

Trend analysis is a purely statistical approach where data are temporally examined. These data sets can be obtained through continuous recording of data of a specific patient. The statistical approach to this is called time series data analysis [30]. In this approach, data sets are assigned a ”time” attribute such that time dependent properties of the data sets can be deduced and analysed. This analysis is important as time patterns and irregularities are critical concerns for the emergence of various diseases. For example, patients often experience immense pain during and after operation and require anesthesia. In normal recovery, the requirement of pain analgesic changes over time. Thus, the analysis of dose delivery information of the analgesic over time on a patient can help predict the variance of a patient pain relief condition [42]. Another application of trend analysis is to follow the population trends of patient populations undergoing a certain treatment for hospital visits, medical costs and lengths of stay of patients [43, 44]. Thus, incurring a trend in the aspects of treatment cost and effectiveness.

1.2.3 Clustering

Simply put, clustering of data is the placing of similar data together in a cluster and dissimilar ones in the others. While clustering can be confused with classification, there is a notable difference among the two. Clustering is an unsupervised learning technique whereas classification is a supervised learning one. Importantly, in clustering the data information about classes is not known. Clustering also does not necessitate the subtle information for the partitioning of the data [45]. A major challenge in this method is that clusters have to be identified first. Typical examples of its application are genomic sequence analysis and genetic expression data analysis [46].

1.2.4 Regression

In regression, data items are analyzed with the motivation of establishing a relationship in the known dependent variable and unknown and independent estimated variable. Statistically, regression is the most effective tool for predicting future patterns [47]. In biomedical research regression correlation coefficients are frequently used to establish a cause and effect relationship. For example, to determine if the patient has high blood pressure and the relationship of the risk of high blood pressure to the weight and age of the patient [48].

1.2.5 Association

Association is the criterion in which the data are examined for the similarities or bonding in which they can be attributed. In examining the data, association rule is very effective. It reveals the correlations and relationships in which the objects are portrayed. Association rules are critical factors in medical marketing, advertising and commodities management [26]. In essence, association rules make it possible for grouping items as per their attributes, then generating rules which can be used conclusively for the data sets. An accurate example is the ranking of hospitals where data mining techniques facilitate the placing of different hospitals according to their performance and other attributes by creating the necessary association on information from various hospitals and then ranking them [23, 49].

1.2.6 Summarization

Using summarization, data can be examined and abstracted to smaller groups or sets of data. The smaller group of data gives the overall description or attributes of the generalized data. The data which are being abstracted can be examined in different ways or perspectives depending on the scope. For instance, this is effectively applied on electronic medical records where the data of the patient population are analysed, categorised based on the data and the insurance providers [28, 50]. By mining the data this way, patterns and regularities of a data set are easily recognised.

1.3 Contributions

The main objective of this thesis is to study the efficiency of ID, GID and IBL mixture models when integrated with the online variational learning algorithm. The contributions

are listed as follows:

- **Medical image segmentation using online variational learning of Finite Inverted Dirichlet Mixture Models approach**

  We propose a finite Inverted Dirichlet mixture model for unsupervised learning using online variational inference. We validate our model on synthetic data and to detect challenging diseases namely brain tumour, lung tuberculosis and skin lesion. This work has been accepted as a book chapter in the book titled *Mixture Models and Applications* [21].

- **Medical image segmentation using online variational learning of Finite Generalized Inverted Dirichlet Mixture Models with feature selection approach**

  An online variational learning algorithm of finite generalized inverted Dirichlet mixture model with feature selection is proposed. Efficiency of proposed model has been evaluated on synthetic data as well as three medical applications for brain tumor detection, skin melanoma detection and computer aid detection (CAD) of malaria. This work has been accepted as a book chapter in the book titled *Artificial Intelligence and data mining in healthcare*.

- **Medical image segmentation using online variational learning of Finite Inverted Beta-Liouville Mixture Models approach**

  We introduce a finite mixture model based on Inverted Beta-Liouville distribution which provides a better fit for the data with online variational learning approach. We evaluated our proposed algorithm on five different biomedical image data sets including optic disc detection and localization in diabetic retinopathy, digital imaging in melanoma lesion detection and segmentation, brain tumour detection, colon cancer detection and computer aid detection (CAD) of Malaria. Furthermore, we compared the proposed algorithm with three other popular algorithms. This work has been submitted to *International Journal of Imaging Systems and Technology* and is under revision.

1.4 Thesis Overview

- ❏ Chapter 1 introduces the concepts of data mining and its use in healthcare and also gives a brief overview of various concepts related to the work proposed. It also conveys clearly our motivations behind the conducted research work.

- ❏ In Chapter 2, we briefly explain the online variational inference approach for inverted Dirichlet mixture models. The efficiency of the model has been evaluated by comparing it to the popular online variational learning of Gaussian Mixture Model (GMM) on synthetic data and to detect challenging diseases namely brain tumour, lung tuberculosis and skin lesion.

- ❏ In chapter 3, we propose online variational learning model for generalized inverted Dirichlet mixture models along with variational feature selection. The model accuracy was tested on different medical data sets using evaluation metrics such as Jaccard Similarity Index, Dice Similarity Coefficient, Adjusted Rand Index (ARI), V-Measure Score and Adjusted Mutual Information (AMI) score.

- ❏ Chapter 4 describes the application of online variational learning to finite inverted Beta-Liouville mixture models. The model has been tested with challenging medical data sets using image segmentation including optic disc detection and localization in diabetic retinopathy.

- ❏ In conclusion, we briefly summarize our contributions.

Chapter 2

Online Variational learning for Finite Inverted Dirichlet Mixture Model

In this chapter, we have examined and analyzed multi-modal medical images by developing an unsupervised machine learning algorithm based on online variational inference for finite Inverted Dirichlet Mixture Model. The prime focus of this chapter is to validate the developed approach on medical images. We do so by implementing the algorithm on both synthetic and real data sets. We test the algorithm's ability to detect challenging diseases namely brain tumour, lung tuberculosis and skin lesion. Extensive comparisons with comparable recent approaches have shown the merits of our proposed model.

2.1 Model Specification

2.1.1 Finite Inverted Dirichlet Mixture Model

The main reason for using finite inverted Dirichlet method is basically to have a flexible distribution for our mixture model. Unlike the Gaussian distribution, it is reasonably flexible and has the property to perform in both symmetric and asymmetric modes. A graphical model for finite inverted Dirichlet mixture model is shown in Figure 2.1. Consider a positive D-dimensional vector that is sampled from a finite inverted Dirichlet mixture model with M components. Hence, the finite mixture of inverted Dirichlet distributions can be defined as:

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{M} \pi_j ID\big(\vec{X}_i \mid \vec{\alpha}_j\big) \tag{2.1}$$

where $\alpha = (\alpha_1, ..., \alpha_M)$ and $\pi = (\pi_1, ..., \pi_M)$ denotes the mixing coefficients along with the constraints that they are positive and sum to one. Also, the term $ID(\vec{X}_i|\vec{\alpha_j})$ hereby represents the j^{th} inverted Dirichlet distribution with the parameter $(\vec{\alpha_j})$ which is defined as [45]:

$$ID\big(\vec{X}_i \mid \vec{\alpha}_j\big) = \frac{\Gamma\big(\sum_{l=1}^{D+1} \alpha_{jl}\big)}{\prod_{l=1}^{D+1} \Gamma\big(\alpha_{jl}\big)} \prod_{l=1}^{D} X_{il}^{\alpha_{jl}-1} \Big(1 + \sum_{l=1}^{D} X_{il}\Big)^{-\sum_{l=1}^{D+1} \alpha_{jl}} \tag{2.2}$$

where, X_{il} is positive for $l = 1, ..., D$ and $\vec{\alpha}_j = \big(\alpha_{j1}, \alpha_{j2}, ...\alpha_{jD+1}\big), \alpha_{jl} > 0$ for $l = 1, ..., D+1$. Mean, variance and co-variance of the inverted Dirichlet distribution are hereby given as under:

$$\mathbb{E}\left[X_l\right] = \frac{\alpha_l}{\alpha_{D+1} - 1} \tag{2.3}$$

$$var\big(X_l\big) = \frac{\alpha_l\big(\alpha_j + \alpha_{D+1} - 1\big)}{\big(\alpha_{D+1} - 1\big)^2\big(\alpha_{D+1} - 2\big)} \tag{2.4}$$

$$cov(X_a, X_b) = \frac{\alpha_a \alpha_b}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2))} \tag{2.5}$$

We introduce an M-dimensional binary random vector $\vec{Z}_i = \{Z_{i1}, ..., Z_{iM}\}$ called the latent variable which is hidden for each of the observed vector X_i. Furthermore, conditional distribution of the Z given the mixing coefficients is as under:

$$p\big(\mathcal{Z} \mid \vec{\pi}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{2.6}$$

Therefore, the conditional distribution of the data set $\mathcal{X}$ can be written as:

$$p\big(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} ID\big(\vec{X}_i \mid \vec{\alpha}_j\big)^{Z_{ij}} \tag{2.7}$$

Assuming that the parameters of the inverted Dirichlet are statistically independent and for every parameter α_{jl}, the Gamma distribution that is adopted to approximate the conjugate prior is given as below :

$$p\big(\alpha_{jl}\big) = \mathcal{G}\big(\alpha_{jl} \mid u_{jl}, \nu_{jl}\big) = \frac{\nu_{jl}^{u_{jl}}}{\Gamma\big(u_{jl}\big)} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{2.8}$$

Here, $\{u_{jl}\}$ and $\{\nu_{jl}\}$ are hyperparameters which have constraint such that $u_{jl} > 0$ and $\nu_{jl} > 0$. Now considering $\vec{\alpha}$ we can write,

$$p(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D} p(\alpha_{jl}) \tag{2.9}$$

The joint distribution of all the random variables can be written as:

$$\begin{aligned} p(\mathcal{X}, \mathcal{Z}, \vec{\alpha} \mid \vec{\pi}) &= p(\mathcal{X} \mid \mathcal{Z}, \vec{\alpha}) p(Z \mid \vec{\pi}) p(\vec{\alpha}) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{M} \left[\pi_j \frac{\Gamma\left(\sum_{l=1}^{D+1} \alpha_{jl}\right)}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} \prod_{l=1}^{D} X_{il}^{\alpha_{jl}-1} \right. \\ &\quad \left. \times \left(1 + \sum_{l=1}^{D+1} X_{il}\right)^{-\sum_{l=1}^{D+1} \alpha_{jl}} \right]^{Z_{ij}} \qquad (2.10) \\ &\quad \times \prod_{j=1}^{M}\prod_{l=1}^{D+1} \frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-v_{jl}\alpha_{jl}} \qquad (2.11) \end{aligned}$$

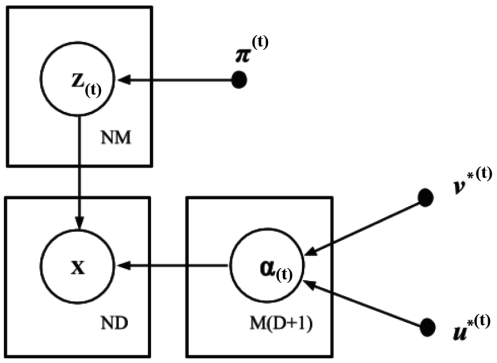

Figure 2.1: Graphical model representation for finite inverted Dirichlet mixture. Symbols in the circle denote the random variables; otherwise, they denote the model parameters.

2.2 Online variational learning for finite inverted Dirichlet mixture model

2.2.1 Variational inference

Variational inference is used to formulate the computation of conditional probability in terms of an optimization problem which is basically deterministic approximation. The main objective of variational inference is to perform an approximation of conditional density of latent variables based on the observed variables. The best choice to find this approximation is by doing optimization. In essence, we make use of a family of densities over the latent variables, which are parameterized by free "variational parameters". Therefore, the task of the optimization is to find the member from this density family i.e., the setting of the parameters, that lie close to the conditional of interest using KL divergence [51]. In order to estimate the parameters of the finite inverted dirchlet mixture model correctly and to select the appropriate number of components for the model , we adopted an online variational approach [52]. For Simplifying the notation, we define $\Theta = \{\mathcal{Z}, \vec{\alpha}\}$. The main purpose of variational learning is to find an approximation $Q(\Theta)$, that approximates $p(\Theta|\mathcal{X}, \vec{\pi})$. To do this, we find the Kullback-Leibler (KL) divergence which is the distance between the the distribution $Q\big(\Theta\big)$ and posterior distribution $p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)$ given by,

$$KL\big(Q \mid\mid P\big) = -\int Q\big(\Theta\big) \ln \left(\frac{p\big(\Theta \mid \mathcal{X}, \vec{\pi}\big)}{Q\big(\Theta\big)}\right) d\Theta \tag{2.12}$$

Modifying this equation we can write

$$KL\big(Q \mid\mid P\big) = \ln\ p\big(\mathcal{X} \mid \vec{\pi}\big) - \mathcal{L}\big(Q\big) \tag{2.13}$$

where, $L(Q)$ is called the variational lower bound , defined as:

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

The KL divergence being a similarity measure follows the conditions $KL\big(Q \mid\mid P\big) \geq 0$ and $KL\big(Q \mid\mid P\big) = 0$ when $Q\big(\Theta\big) = p\big(\Theta \mid \mathcal{X}\big)$. From $\big(2.13\big)$ we can say $\mathcal{L}\big(Q\big)$ is the lower bound of $p\big(\mathcal{X} \mid \vec{\pi}\big)$. We maximize the lower bound which means we are minimizing the KL divergence and hence approximating the true posterior distribution. However, the true posterior distribution cannot be used directly for variational inference as

it is computationally intractable. Therefore, for this reason we use the method of mean-field approximation for our algorithm [53][54][55] by which we factorize $Q\big(\Theta\big)$ into disjoint tractable distributions as below

$$Q\big(\Theta\big) = Q\big(\mathcal{Z}\big)Q\big(\vec{\alpha}\big) \tag{2.15}$$

To maximize the lower bound $L(Q)$, we are supposed to make a variational optimization of $L(Q)$ with respect to each factor. The variational solution for a specific parameter $Q_k\big(\Theta_k\big)$ is

$$Q_k\big(\Theta_k\big) = \frac{exp\big\langle \ln\ p\big(\mathcal{X}, \Theta\big)\big\rangle_{\neq k}}{\int exp\big\langle \ln\ p\big(\mathcal{X}, \Theta\big)\big\rangle_{\neq k} d\Theta} \tag{2.16}$$

where $\big\langle . \big\rangle_{\neq k}$ is the expectation with respect to all the parameters other than Θ_k.

We hereby can obtain the following optimal variational solutions for the finite inverted Dirichlet mixture model (derived in Appendix A.1 and Appendix A.2)

$$Q\big(\mathcal{Z}\big) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{2.17}$$

$$Q(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D+1} G(\alpha_{jl}|u_{jl}^{*}, \nu_{jl}^{*}) \tag{2.18}$$

where,

$$r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M}\rho_{ij}} \tag{2.19}$$

$$\rho_{ij} = exp\Bigg\{ \ln\pi_j + \tilde{R}_j + \sum_{l=1}^{D}\big(\overline{\alpha}_{jl} - 1\big)\ln X_{il} - \sum_{l=1}^{D+1}\overline{\alpha}_{jl}\ \ln\big(1 + \sum_{l=1}^{D} X_{il}\big)\Bigg\} \tag{2.20}$$

$$
\begin{aligned}
\tilde{R}_j = \ln \frac{\Gamma\big(\sum_{l=1}^{D+1} \overline{\alpha}_{jl}\big)}{\prod_{l=1}^{D+1} \Gamma\big(\overline{\alpha}_{jl}\big)} \\
&+ \sum_{l=1}^{D+1} \overline{\alpha}_{jl} \left[\psi\left(\sum_{l=1}^{D+1} \overline{\alpha}_{jl} \right) - \psi\big(\overline{\alpha}_{jl}\big) \right] \Big[\big\langle \ln \alpha_{jl} \big\rangle - \ln \overline{\alpha}_{jl} \Big] \\
&+ \frac{1}{2} \sum_{l=1}^{D+1} \overline{\alpha}_{jl}^2 \left[\psi'\left(\sum_{l=1}^{D+1} \overline{\alpha}_{jl} \right) - \psi'\big(\overline{\alpha}_{jl}\big) \right] - \Big\langle \big(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl} \big)^2 \Big\rangle \\
&+ \frac{1}{2} \sum_{a=1}^{D+1} \sum_{b=1}^{D+1} \overline{\alpha}_{ja}\, \overline{\alpha}_{jb} \left[\psi'\left(\sum_{l=1}^{D+1} \overline{\alpha}_{jl} \right) \Big(\big\langle \ln \overline{\alpha}_{ja} \big\rangle - \ln \overline{\alpha}_{ja} \Big) \right. \\
&\left. \times \Big(\big\langle \ln \overline{\alpha}_{jb} \big\rangle - \ln \overline{\alpha}_{jb} \Big) \right]
\end{aligned} \tag{2.21}
$$

The estimation equations for u_{jl}^* and v_{jl}^* are given by (derived in Appendix A.2)

$$
\begin{aligned}
u_{jl}^* = u_{jl} + \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \overline{\alpha}_{jl} \left[\psi\left(\sum_{s=1}^{D+1} \overline{\alpha}_{js} \right) - \psi\big(\overline{\alpha}_{jl}\big) \right. \\
\left. + \sum_{s \neq l}^{D+1} \psi'\left(\sum_{ls=1}^{D+1} \overline{\alpha}_{js} \right) \times \overline{\alpha}_{js} \Big(\big\langle \ln \alpha_{js} \big\rangle - \ln \overline{\alpha}_{js} \Big) \right]
\end{aligned} \tag{2.22}
$$

$$
v_{jl}^* = v_{jl} - \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \left[\ln X_{il} - \ln \left(1 + \sum_{l=1}^{D+1} X_{il} \right) \right] \tag{2.23}
$$

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

$$
\big\langle Z_{ij} \big\rangle = r_{ij} \tag{2.24}
$$

$$
\overline{\alpha}_{jl} = \big\langle \alpha_{jl} \big\rangle = \frac{u_{jl}^*}{\nu_{jl}^*}, \quad \big\langle \ln \alpha_{jl} \big\rangle = \psi\big(u_{jl}^*\big) - \ln \nu_{jl}^* \tag{2.25}
$$

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

We therefore maximize the variational lower bound $L(Q)$ to estimate the coefficient $\vec{\pi}$ which is treated as parameter for mixture model. The derivative of this lower bound with

respect to $\vec{\pi}$ (derived in Appendix A.3) is given as under:

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

Therefore, for the variational learning of inverted Dirichlet mixture model, the value of the lower bound is calculated as:

$$\begin{aligned} L(Q) &= \sum_{z}\int Q(Z,\vec{\alpha}) ln\{\frac{p(\chi, Z, \vec{\alpha}|\vec{\pi})}{Q(Z,\vec{\alpha})}\} d\alpha \\ &=< lnp(\chi|Z,\vec{\alpha}) > + < lnp(Z|\vec{\pi}) > + < lnp(\vec{\alpha}) > \\ &- < lnQ(Z) > - < lnQ(\vec{\alpha}) > \end{aligned} \tag{2.28}$$

2.2.2 Online variational inference

In this section, we present an online variational inference algorithm for finite inverted Dirichlet mixture models. In this algorithm we treat variational inference as a natural gradient which is the inverse of the Riemannian metric multiplied by the gradient [56]. We do this as it helps to achieve optimal convergence which allows to have faster online inference.

Online learning is when the data become available in a sequence and later the previous data are used to as a reference to update the best predictor for the new incoming data at each step since the data is continuously arriving in online fashion. It is different from batch learning variational technique, in which we know the best predictor by working on the entire data set at the same time. Online learning is being commonly used in many areas where it is completely infeasible to train the entire data set at once since the data set is too large to be trained altogether. Online learning is also extensively useful in areas such as stock price prediction where it is important to adapt to the new patterns in the data or even when the data itself is generated as a function of time. In such a case when the data are continuously arriving in an online fashion, we have to estimate the variational lower bound to a fixed amount of data which is N. Considering this, the value expected from the model evidence $p(X)$ for a data with finite size can be derived as [31]:

$$\left\langle \ln p(X) \right\rangle_{\phi} = \int \phi(X) \ln \left(\int p(X|\Theta) p(\Theta) d(\Theta) \right) dx \tag{2.29}$$

where $\phi(X)$ represents the probability distribution which is unknown for the data observed. Thus, the corresponding expected variational lower bound can be computed using [31]:

$$\begin{aligned}\left\langle \mathcal{L}(Q)\right\rangle_{\phi} &= \left\langle \sum_{\mathcal{Z}} \int Q(\alpha)Q(\mathcal{Z}) \ln\left[\frac{p(X,\mathcal{Z}|\alpha)p(\alpha)}{Q(\alpha)Q(\mathcal{Z})}\right] d\alpha \right\rangle_{\phi} \\ &= N \int Q(\alpha) d\alpha \left\langle \sum_{Z} Q(Z) \ln\left[\frac{p(X,Z|\alpha)}{Q(Z)}\right] \right\rangle_{\phi} \\ &+ \int Q(\alpha) \ln\left[\frac{p(\alpha)}{Q(\alpha)}\right] d\alpha \end{aligned} \tag{2.30}$$

We consider t as the actual amount of data observed thus for the observed data the current lower bound can be estimated by [31]

$$\begin{aligned}\mathcal{L}^{(t)}(Q) =& \frac{N}{t} \sum_{i=1}^{t} \int Q(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \sum_{\mathbf{Z}_i} Q\left(\mathbf{Z}_i\right) \ln\left[\frac{p\left(\mathbf{X}_i, \mathbf{Z}_i|\boldsymbol{\alpha}\right)}{Q\left(\mathbf{Z}_i\right)}\right] \\ &+ \int Q(\boldsymbol{\alpha}) \ln\left[\frac{p(\boldsymbol{\alpha})}{Q(\boldsymbol{\alpha})}\right] \mathrm{d}\boldsymbol{\alpha} \end{aligned} \tag{2.31}$$

We realise that while N remains fixed, t increases over time. The main reason for this is the fact that the principal objective of the proposed online algorithm is the expected log evidence computed for a fixed amount of data. Even if there is an increase in the observed data, the algorithm basically computes the same quantity. Now relating this to the context, the former observed data is then used to improve the quality of estimation of the expected variational lower bound in equation (2.30). This inherently approximates the resulting log evidence as it does not have any previous knowledge of the former observed data.

With respect to the expectation values we saw in previous section, equation (2.25) and (2.26) for $i = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, D+1$ get modified to the below equations as the data is getting updated in online fashion.

$$\overline{\alpha}_{jl} = \left\langle \alpha_{jl} \right\rangle = \frac{u_{jl}^{(t-1)}}{\nu_{jl}^{(t-1)}}, \quad \left\langle \ln \alpha_{jl} \right\rangle = \psi\big(u_{jl}^{(t-1)}\big) - \ln \nu_{jl}^{(t-1)} \tag{2.32}$$

$$\left\langle \big(\ln \alpha_{jl} - \ln \overline{\alpha}_{jl}\big)^2 \right\rangle = \Big[\psi\big(u_{jl}^{(t-1)}\big) - \ln u_{jl}^{(t-1)}\Big]^2 + \psi'\big(u_{jl}^{(t-1)}\big) \tag{2.33}$$

The fundamental concept of this online algorithm is to enable successful maximization of the present variational lower bound in equation (2.31). Assuming that the observed data set exists in the form $\{X_1, \ldots X_{t-1}\}$. For every new observation X_t , we mainly perform maximization of the present $\mathcal{L}^{(t)}(Q)$ with respect to $Q\big(\mathcal{Z}_t\big)$ while $Q(\alpha)$ is set to $Q^{(t-1)}(\alpha)$ and π_j is set to $\pi_j^{(t-1)}$. Hence, the variational solution can be computed using:

$$Q\big(\mathcal{Z}_t\big) = \prod_{j=1}^{M} r_{tj}^{Z_{tj}} \tag{2.34}$$

where, ρ_{ij}

$$\rho_{ij} = \exp[ln\pi_j^{(t-1)} + R_j + \sum_{l=1}^{D}(\vec{\alpha_{jl}} - 1)lnX_{il} - (\sum_{l=1}^{D+1}\vec{\alpha_{jl}})ln(1 + \sum_{l=1}^{D} X_{il})] \tag{2.35}$$

where R_j is given by

$$\begin{aligned}
\tilde{R}_j = \ln & \frac{\Gamma\big(\sum_{l=1}^{D+1}\overline{\alpha}_{jl}\big)}{\prod_{l=1}^{D+1}\Gamma\big(\overline{\alpha}_{jl}\big)} \\
& + \sum_{l=1}^{D+1}\overline{\alpha}_{jl}\left[\psi\left(\sum_{l=1}^{D+1}\overline{\alpha}_{jl}\right) - \psi\big(\overline{\alpha}_{jl}\big)\right]\Big[\big\langle \ln\alpha_{jl}\big\rangle - \ln\overline{\alpha}_{jl}\Big] \\
& + \frac{1}{2}\sum_{l=1}^{D+1}\overline{\alpha}_{jl}^2\left[\psi'\left(\sum_{l=1}^{D+1}\overline{\alpha}_{jl}\right) - \psi'\big(\overline{\alpha}_{jl}\big)\right] - \Big\langle\big(\ln\alpha_{jl} - \ln\overline{\alpha}_{jl}\big)^2\Big\rangle \\
& + \frac{1}{2}\sum_{a=1}^{D+1}\sum_{b=1}^{D+1}\overline{\alpha}_{ja}\,\overline{\alpha}_{jb}\left[\psi'\left(\sum_{l=1}^{D+1}\overline{\alpha}_{jl}\right)\Big(\big\langle\ln\overline{\alpha}_{ja}\big\rangle - \ln\overline{\alpha}_{ja}\Big)\right. \\
& \left.\times\Big(\big\langle\ln\overline{\alpha}_{jb}\big\rangle - \ln\overline{\alpha}_{jb}\Big)\right]
\end{aligned} \tag{2.36}$$

Later, we maximize the lower bound $\mathcal{L}^{(t)}(Q)$ with respect to $Q^{(t)}(\alpha)$ and $\pi_j^{(t)}$ while $Q\big(\mathcal{Z}_t\big)$ is fixed. As mentioned before, here we consider variational inference as a natural gradient method. Therefore, the coefficient matrix for the posterior parameter distribution gets cancelled since the natural gradient of a parameter is obtained by multiplying the gradient by the inverse of Riemannian metric. Therefore, the natural gradients for Δu_{js}, $\Delta\nu_{js}$ for $j = 1, 2, ..., M$ and $s = 1, 2, ..., D + 1$ are

$$\Delta u_{js} = \vec{\alpha_{js}}[\psi\big(\sum_{l=1}^{D+1}\vec{\alpha_{jl}}\big) - \psi\big(\vec{\alpha_{js}}\big) + \psi'\big(\sum_{l=1}^{D+1}\big)\sum_{l\neq s}^{D+1}\vec{\alpha_{jl}}(< ln\alpha_{jl} > -ln\vec{\alpha_{jl}})]\sum_{i=1}^{N} r_{ij} \tag{2.37}$$

$$\Delta\nu_{js} = -\sum_{i=1}^{N} r_{ij}[lnX_{is} - ln(1+\sum_{l=1}^{D} X_{il})] \tag{2.38}$$

Thus, the variational solution to $Q^{(t)}(\alpha)$ is given by

$$Q^{(t)}(\alpha) = \prod_{j=1}^{M}\prod_{l=1}^{D+1} G(\alpha^*_{jl}|u^*_{jl},\nu^*_{jl}) \tag{2.39}$$

Therefore, we update the hyper parameters and optimal variational parameters as

$$u^{(t)}_{jl} = u^{(t-1)}_{jl} + \rho_t \Delta u_{jl} \tag{2.40}$$

$$\nu^{(t)}_{jl} = \nu^{(t-1)}_{jl} + \rho_t \Delta \nu_{jl} \tag{2.41}$$

where ρ_t is learning rate in which $epsilon \in$ (0,1) and $\eta_o \geq 0$ is defined as

$$\rho_t = (\eta_o + t)^{-\epsilon} \tag{2.42}$$

The function of the learning rate here is adopted from [57] and is used to forget the earlier inaccurate estimation effects that contributed to the lower bound and expedite the convergence of the learning process. Online learning embraces the fact that learning environments can (and do) change from second to second. The mixing coefficient $\pi^{(t)}_{jl}$ is given by

$$\pi^{(t)}_{jl} = \pi^{(t-1)}_{jl} + \rho_t \Delta \pi_{jl} \tag{2.43}$$

where $\Delta\pi_j$ is

$$\Delta\pi_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} - \pi^{(t-1)}_{j} \tag{2.44}$$

The varitional lower bound in case of online variational inference does not always increase where as in batch variational it does because in case of online learning a new contribution is always added to the lowerbound for each new observation. It is very important

to choose the hyper parameters and the learning rate accurately since it might affect the convergence of the model.

Algorithm 1 Online Variational learning of the finite inverted Dirichlet mixture model.

1. Choose the initial number of components M.
2. Initialize the value of hyper-parameters values for u_{jl} and ν_{jl}.
3. Using K-means algorithm, initialize the value of r_{ij}.
4. **for** $t = 1 \rightarrow N$ **do**
   - i The variational E-step:
   - ii Update the variational solutions for $Q(Z_t)$ using r_{ij}
   - iii The variational M-step:
   - iv Compute the learning rate $\rho_t = (\eta_o + t)^{-\epsilon}$
   - v Calculate the natural gradients Δu_{js}, $\Delta \nu_{js}$ and $\Delta \pi_j$ using (2.37) , (2.38) and (2.44) respectively.
   - vi Update the variational solution for $Q^{(t)}(\alpha)$ and the mixing coefficient $\pi_{jl}^{(t)}$ through (2.39) and (2.43)
   - vii Repeat the variational E-step and M-step until new data is observed.
5. **end for**

2.3 Experimental results

In order to evaluate the performance of our proposed algorithm we first validate it on synthetic data sets of varied sizes. Once the algorithm is validated, we further apply it on real world medical image data sets which are available with ground truth to perform segmentation and analysis of diseases. In our case, we performed medical image segmentation on three data sets of different diseases and different medical image testing techniques. We applied the algorithm to detect brain tumour, skin lesion and tuberculosis. Furthermore, we have used three different formats of images to test the applicability of the algorithm on varied output formats, namely, MRI scans, normal photographs and X-ray images.

In order to have an insight on the accuracy of our algorithm we further compared it to the implementation of online variational inference of finite Gaussian mixture model on the data sets. We chose online variational inference of finite Gaussian mixture model as the comparison algorithm since Gaussian mixtures are widely applied in medical applications.

Image segmentation is a key challenge in image analysis. In medical imaging, it is a particularly difficult challenge due to high variability in the image data sets. This variability arises due to two reasons. One, each human itself has variability in the anatomy of the organ or tissue. Second, there is an additional technical variability introduced to the images due to the different modalities (e.g., MRI, PET scans, CT scans etc.) by which the image is created.

Let's say we have an input observed dataframe X which contains N pixels such that X= $\{X_1, \ldots X_N\}$. Each pixel is modelled as a mixture of M inverted Dirichlet distributions :

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\alpha}\big) = \sum_{j=1}^{M} \pi_j ID\big(\vec{X}_i \mid \vec{\alpha}_j\big) \tag{2.45}$$

where X_i is the pixel intensity value. In all our experiments, we initialize the number of components M to 15. The parameters of the ϵ and η_o learning rate are set to 0.1 and 64 respectively. The accuracy of the algorithm was verified by comparison with the ground truth that were available for each data sets. According to our experiments, a good choice of the initial values of the hyper parameters u_{jl} and ν_{jl} are discovered to be 1 and 0.01 respectively. We can thus detect the optimal number of the components M by eliminating the components with the small mixing coefficients close to 0.

2.3.1 Synthetic data

The goal of using synthetic data is to investigate the accuracy of the online variational approach for both parameter estimation and model selection. Therefore, we first tested the model accuracy on synthetic data sets. These data sets consisted of different data sizes, namely, 300, 400, 600, 800 and 1000. The effectiveness of the algorithm was tested by estimating the mixture parameters. Table (2.1) represents a comparison of the estimation performed by online variational learning of inverted Dirichlet mixture model versus the

real parameters. It is noted that our algorithm can determine mixing coefficient parameters ($\hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \hat{\alpha}_{j3}$ and $\hat{\pi}_j$) close to the real data parameters ($\alpha_{j1}, \alpha_{j2}, \alpha_{j3}$ and π_j).

There are different ways that can be used for estimation of the number of components. In our case, once the algorithm reached convergence, we removed the components with very small (less than 10^{-5}) mixing coefficients in each data set.

Data set	N_j	j	α_{j1}	α_{j2}	α_{j3}	π_j	$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\alpha}_{j3}$	$\hat{\pi}_j$
S1	100	1	7	25	79	0.33	7.43	25.62	80.28	0.33
$\big(N = 300\big)$	100	2	11	32	63	0.34	10.41	31.81	62.97	0.35
	100	3	22	45	51	0.33	22.37	43.99	51.09	0.33
S2	200	1	7	25	79	0.50	6.89	25.82	81.9	0.49
$\big(N = 400\big)$	200	2	11	32	63	0.50	11.52	33.9	67.36	0.51
S3	200	1	7	25	79	0.33	7.56	25.62	79.3	0.33
$\big(N = 600\big)$	200	2	11	32	63	0.33	11.26	31.55	62.78	0.33
	200	3	22	45	51	0.33	22.42	46.66	51.55	0.33
S4	200	1	7	25	79	0.25	7.5	25.53	82.88	0.24
$\big(N = 800\big)$	200	2	11	32	63	0.25	11.32	33.21	68.4	0.25
	400	3	22	45	51	0.50	21.2	44.55	50.87	0.5
S5	200	1	7	25	79	0.25	6.98	24.95	78.76	0.24
$\big(N = 800\big)$	200	2	11	32	63	0.25	10.21	29.43	60.29	0.25
	200	3	22	45	51	0.25	22.11	45.11	50.85	0.27
	200	4	28	83	90	0.25	28.75	85.4	93.05	0.23
S6	200	1	7	25	79	0.20	7.32	24.95	76.64	0.18
$\big(N = 1000\big)$	200	2	11	32	63	0.20	11.86	34.79	68.16	0.21
	200	3	22	45	51	0.20	23	46.1	51.59	0.22
	200	4	28	83	90	0.20	28.71	83.81	88.55	0.23
	200	5	40	3	56	0.20	37.94	2.95	55.05	0.2

Table 2.1: Real and estimated parameters of different data sets. N denotes the total number of data points, N_j denotes the number of data points in the cluster j. $\alpha_{j1}, \alpha_{j2}, \alpha_{j3}$ and π_j are the real parameters and $\hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \hat{\alpha}_{j3}$ and $\hat{\pi}_j$ are the parameters estimated by our proposed algorithm.

2.3.2 Medical image data sets

After validating the algorithm on synthetic data sets, we applied it on three biomedical image data sets. These data sets were used to detect three different disease morphology's which were created using three different imaging techniques. These data sets were MRI

scan of brain tumors, X-ray scans of lung tuberculosis and normal png format pictures of skin lesions. We observed that our algorithm could detect the morphological and structural anomalies similar to the ground truth data. We used 25 images in each case and compared the results of our proposed algorithm with online variational learning for Gaussian mixture model (OVGMM) to determine the model performance.

Brain tumour detection

Gliaomas or brain tumor are the most prominent brain malignancies which exhibit varying degrees of aggressiveness, prognosis and inherent variability in the MRI image representation. Due to the heterogeneous nature of the brain anatomy, the MRI image segmentation and tumor detection is a highly challenging task [58]. For this, the brain tumor data set was obtained from BRATS2015[1] [1, 2]. The data set consists of four MRI sequence images for each patient. The MRI sequence images were Fluid Attenuation Inversion Recovery (FLAIR), T1c, T1p and T2 which all stand for images which are weighted with respect to the relaxation time of protons in the body tissue during the scanning. FLAIR is widely applied to detect clinical malformations related to diseases like Multiple sclerosis (MS), Hemorrhages, Meningitis etc [59]. In our experiment, we used the available FLAIR images for image segmentation and brain tumor detection.

The resulting accuracy of brain MRI segmentation was measured using Jaccard and Dice metrics. This is illustrated in Figure 2.2 and the mean and standard deviation are illustrated in Figure 2.3.The Jaccard and Dice for the BRATS2015[2] data set were significantly greater for our proposed algorithm than the online variational Gaussian Mixture Model helping us conclude that it can be useful to detect tumours. The mean was 0.5 greater than the compared algorithm and the standard deviation was comparatively less showing the robustness of our model.

[1]https://www.smir.ch/BRATS/Start2015

[2]https://www.smir.ch/BRATS/Start2015

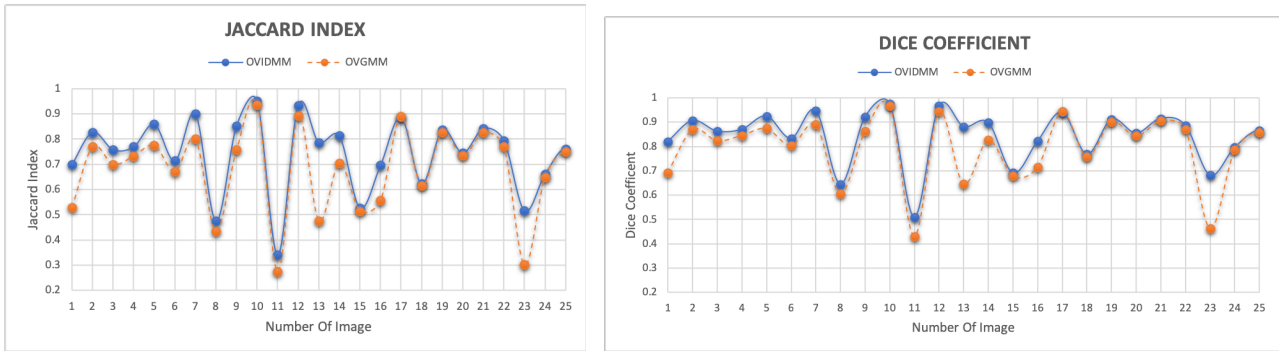

Figure 2.2: Results using Jaccard and dice evaluation metrics for brain tumour detection

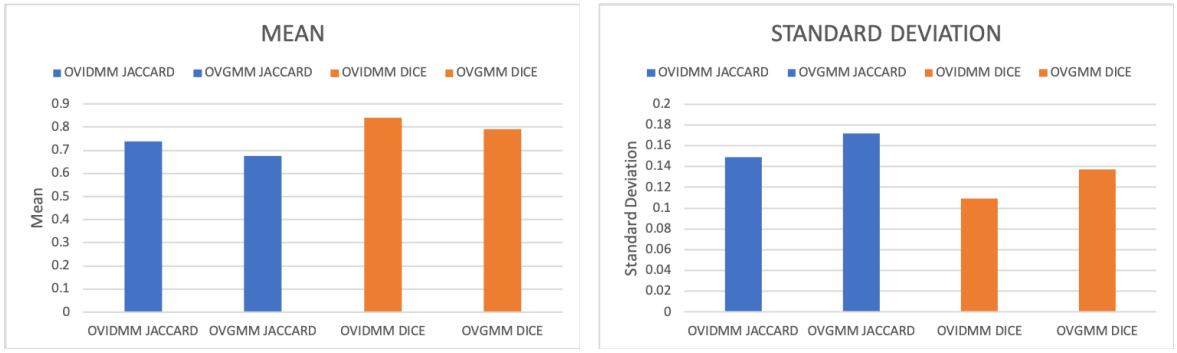

Figure 2.3: Mean and standard deviation results for brain tumour detection

Representative segmentation results after running our proposed algorithm are depicted in Figure 2.4 where the three clusters generated by the algorithm are depicted against the MRI image. The last image in the panel is the best prediction made by the algorithm and it is seen that the algorithm is able to identify the brain glioma. Further, the post processing images are depicted in Figure 2.5 where, the last image in the panel is the post processed ground truth image. The predicted image by the algorithm was compared against the ground truth. We are able to visibly see the similarities of the detection by the algorithm versus an experts opinion.

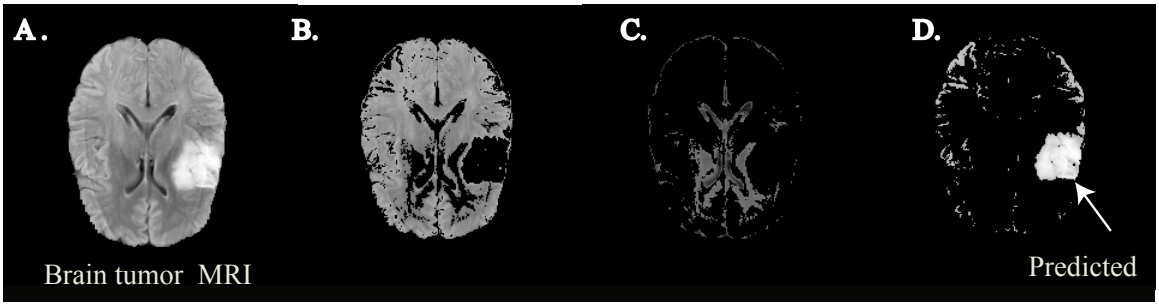

Figure 2.4: Best segmented brain MRI images : A. Input image, B. 3rd Cluster, C. 6th Cluster, D. 7th Cluster

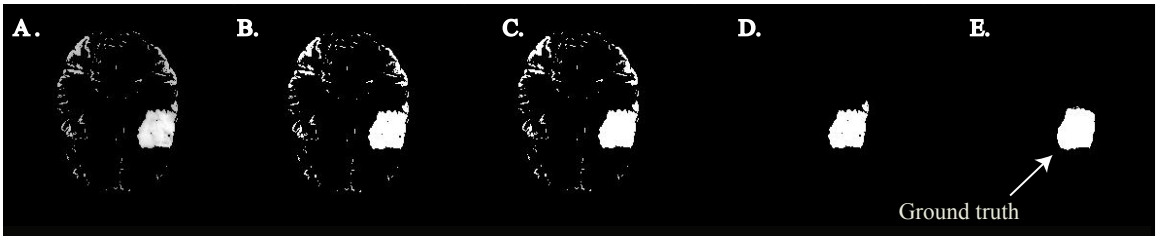

Figure 2.5: Segmented brain MRI images after post processing: A.Clustered image, B. Binary image, C. Clustered after filling holes, D. Processed clustered image and E. Ground truth image. The data set was taken from BRATS database [1, 2] where the ground truth data was available.

Skin lesion diagnosis

Similarly to brain gliomas, skin melanomas are also difficult glioma to detect. Specially because the naked eye is not able to differentiate between the malignant and benign skin melanoma [60]. Therefore digital imaging and lesion detection with identification can help increase the efficiency in the detection and treatment [61]. Furthermore, since skin is the largest organ of the body and highly visible, taking photos of the melanomas from smart phones would add convenience in the process. However analysis of smart phone medical images is also a challenging task due to the heterogeneity [62, 63]. For this reason, the data used for assessing the performance of the proposed algorithm was done on the photos of skin lesion obtained from International Skin Imaging Collaboration[3]. The data set consists of images of skin melanoma of patients.

[3]https://isic-archive.com/api/v1

The accuracy of the result obtained from skin image segmentation is measured by Jaccard and Dice metrics as illustrated in Figure 2.6 and the mean and standard deviation are shown in Figure 2.7 by comparing the proposed algorithm with online variational finite Gaussian mixture model. The Jaccard index and Dice coefficient for the data set were significantly greater for our proposed algorithm since both the values for each image were above 0.85. The mean was 0.7 greater than the compared algorithm and the standard deviation was 0.05 less for our algorithm proving the robustness of our algorithm. This demonstrates the accuracy of our model for predicting skin lesions.

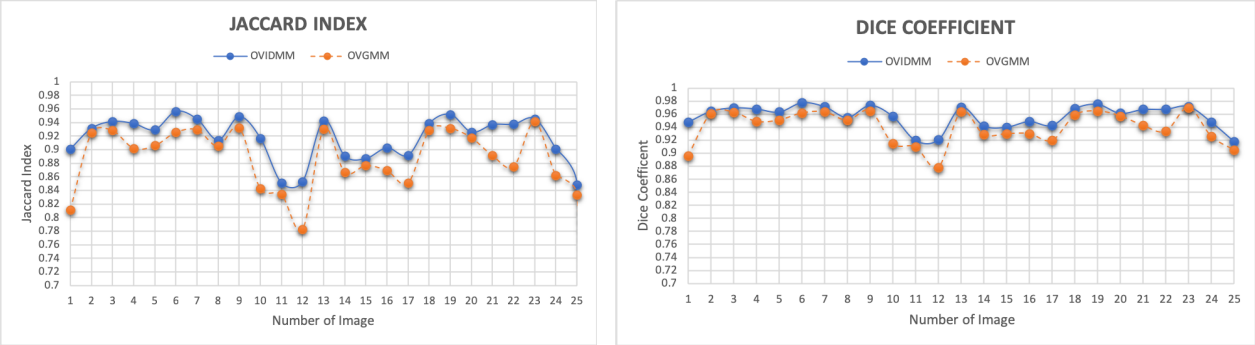

Figure 2.6: Results using Jaccard and Dice evaluation metrics for skin lesion diagnosis

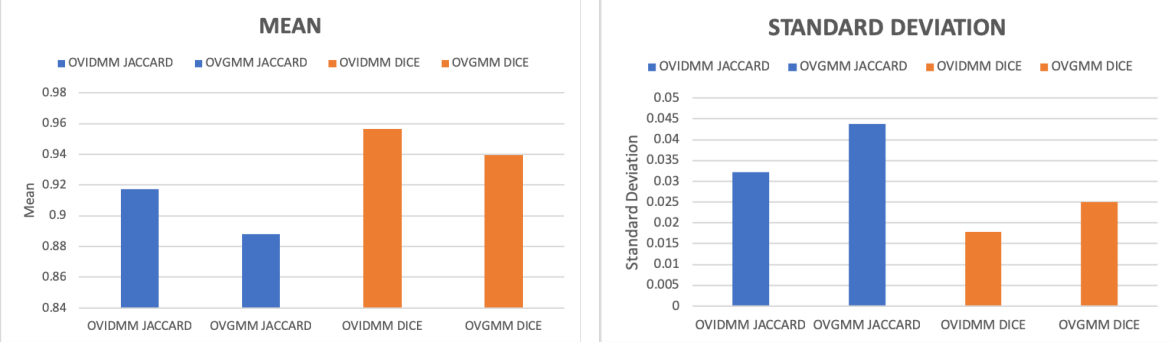

Figure 2.7: Mean and standard deviation results for skin lesion diagnosis

Figure 2.8 shows a representative image of skin melanoma from the ISIC database (left panel, first photo) and the best segmented and detected melanoma by the algorithm can be seen at the end of the panel in the figure. In this case the algorithm was able to detect 14 clusters. Figure 2.9 displays a representative skin melanoma image achieved after post processing for the ground truth in order to compare it with the algorithm.

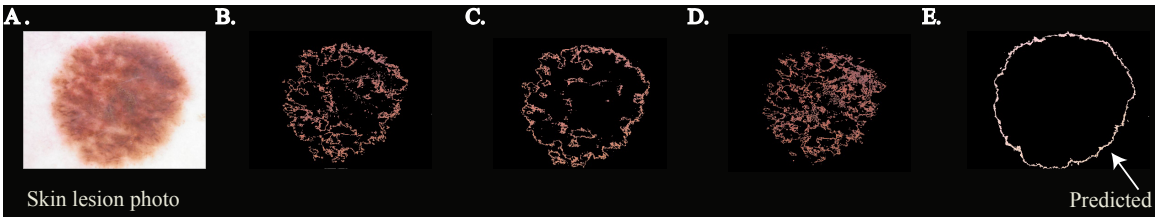

Figure 2.8: Best Segmented Skin Lesion Images: A. Input image, B. 0^{th} Cluster, C. 9^{th} Cluster, D. 14^{th} Cluster E. 10^{th} Cluster

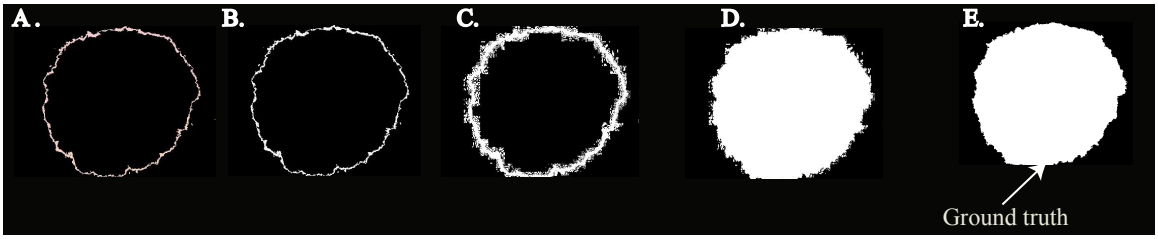

Figure 2.9: Segmented Skin lesion images after post processing: A.Clustered image, B. Greyscale image, C. Binary image, D. Binary image after filling holes, and E. Ground truth image. The data set was taken from ICIS database where the ground truth data was available.

Lung tuberculosis detection

Tuberculosis is caused by *Mycobacterium tuberculosis* which majorly infects the lung but can spread rapidly through the body [64]. X-Ray is currently the most common diagnostic tool used to detect tuberculosis. However, a lot of time the infection goes undetected due to the high intrinsic noise in the X-Ray measurements [65]. Besides, in a low resource setup X-Ray interpretations are performed by non-experts [66]. Here, a digital analysis of detection can lend to computer aided decision support. Therefore, the third data set used for this analysis is an X-Ray image selected from collection of data compiled by National Library of Medicine in collaboration with the Department of Health and Human Services, Montgomery County, Maryland, USA [67, 68]. The sample set is composed of 58 cases with manifestation of tuberculosis and 80 normal cases. Each image is gray-scale with a spatial resolution of 4020 x 4892, or 4892 x 4020. We performed our algorithm on 25 images and on cases where Tuberculosis was detected. It is to be noted that we compared only the right mask of the lung for the algorithm predictions as the ground truth was available for that.

The accuracy obtained by performing lung segmentation is given by Jaccard and Dice metrics as illustrated in Figure 2.10 and the mean and standard deviation are shown in Figure 2.11. The Jaccard and Dice for this data set were significantly higher for our proposed algorithm since each image had a considerable difference in the value from the online variational learning of finite Gaussian Mixture Model. The mean was 0.11 greater than the compared algorithm and the standard deviation was comparatively less for our algorithm showing the strength of our model.

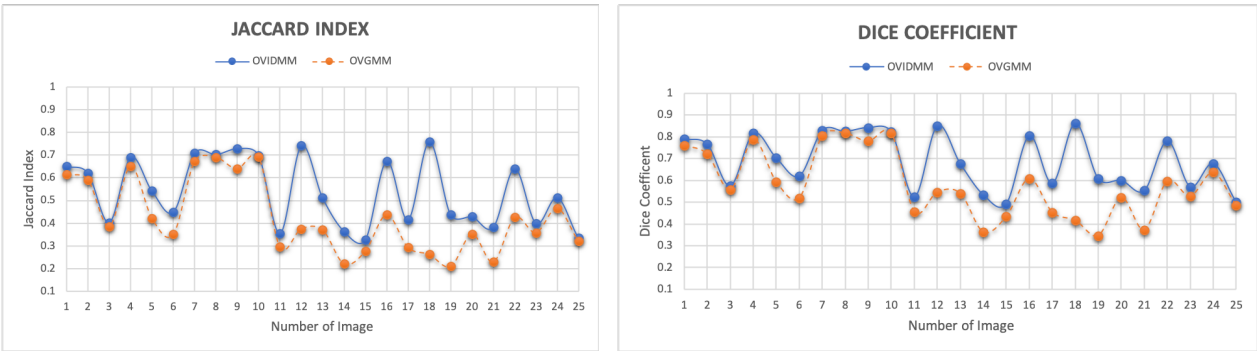

Figure 2.10: Results using Jaccard and Dice evaluation metrics for lung tuberculosis detection

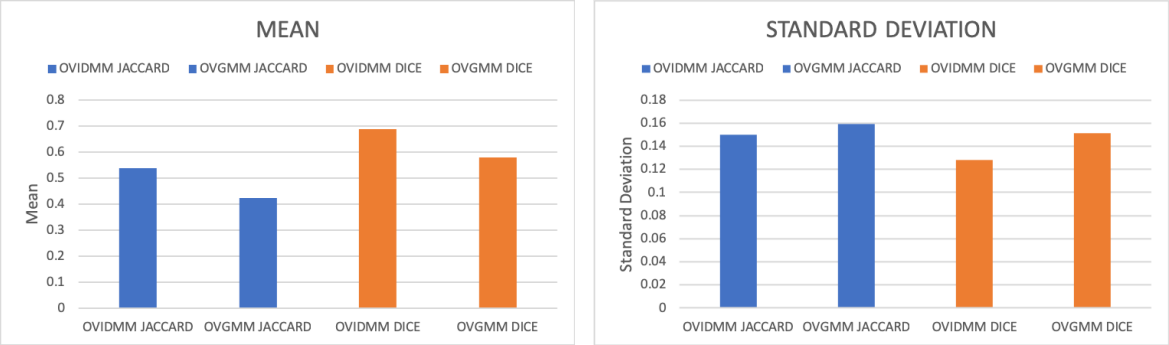

Figure 2.11: Mean and standard deviation results for lung tuberculosis detection

Figure 2.12 is a representative image of the prediction algorithm where the top 4 clusters are depicted. In the panel the first image is that of the X-Ray and the last image is of the best predicted tuberculosis image by the algorithm. There were 14 clusters generated by the algorithm which are not shown here. Figure 2.13 depicts the images of the same lung X-Ray segmentation after post processing on segmentation. It can be clearly seen in the

last images of Figure 2.12 (predicted) and Figure 2.13 (ground truth) that the algorithm is able to capture the similar segment of tuberculosis in the right lung.

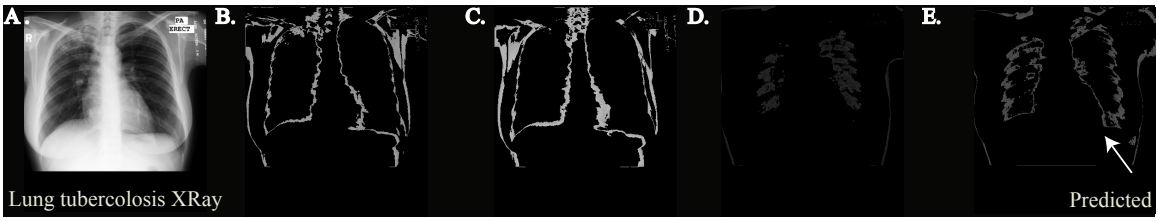

Figure 2.12: Best segmented Lung images : A. Input image, B. 10^{th} Cluster, C. 7^{th} Cluster, D. 4^{th} Cluster,E. 0^{th} Cluster

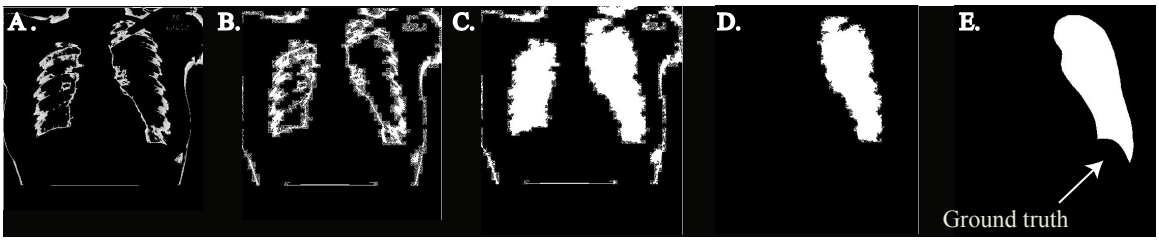

Figure 2.13: Lung X-ray after post processing: A.Clustered image, B. Binary image, C. Clustered after filling holes, D. Processed cluster and E. Ground truth image. The data set was taken from Montgomery County - Chest X-ray Database provided by national library of medicine where the ground truth data was available.

Chapter 3

Online Variational learning using Finite Generalized Inverted Dirichlet Mixture Model with Feature Selection

In this chapter, we propose a statistical framework for online variational learning of finite generalized inverted Dirichlet (GID) mixture model for clustering medical images data by simultaneously using feature selection and image segmentation. The model allows one to adjust the mixture model parameters, number of components and features weights to tackle the challenge of over fitting. The algorithm in this study has been evaluated on synthetic data as well as three medical applications for brain tumor detection, skin melanoma detection and computer aid detection (CAD) of malaria.

3.1 Model specification

The most significant reason to consider generalized inverted Dirichlet distribution as a standard one in our mixture model is its ability to generate models specified to positive vectors and its more general covariance structure. The GID has several interesting mathematical properties which allow for instance, the representation of GID samples in a transformed space in which features are independent and follow inverted Beta distributions [69]. We consider a set $\mathcal{Y}$ of N D-dimensional positive vectors, such that $\mathcal{Y} = (Y_1, Y_2, ..., Y_N)$ and M indicates the number of various clusters [70]. We suppose that $\mathcal{Y}$ is managed by a mixture

of GID distributions $p\big(\mathcal{Y}_i \mid \vec{\pi}, \vec{\alpha}, \vec{\beta}, \big)$ [71] as

$$p\big(\mathcal{Y}_i \mid \vec{\pi}, \vec{\alpha}, \vec{\beta}\big) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} \frac{\Gamma\big(\alpha_{jd} + \beta_{jd}\big)}{\Gamma\big(\alpha_{jd}\big)\Gamma\big(\beta_{jd}\big)} \frac{\mathcal{Y}_{id}^{\alpha_{jd}-1}}{\big(1 + \sum_{l=1}^{d} \mathcal{Y}_{il}\big)^{\gamma_{jd}}} \tag{3.1}$$

where $\vec{\alpha} = \big\{\alpha_1, \alpha_2, ...\alpha_M\big\}$, with $\vec{\alpha}_j = \big\{\alpha_{j1}, \alpha_{j2}, ...\alpha_{jD}\big\}$, j = 1,.., M and $\vec{\beta} = \big\{\beta_1, \beta_2, ...\beta_M\big\}$, with $\vec{\beta}_j = \big\{\beta_{j1}, \beta_{j2}, ...i\beta_{jD}\big\}$, j = 1,.., M. $\vec{\pi} = \big\{\pi_1, \pi_2, ...\pi_M\big\}$, are the mixing weights, such that $\sum_{j=1}^{M} \vec{\pi_j} = 1$. We define $\vec{\gamma_{jd}}$ such that $\vec{\gamma}_{jd} = \vec{\beta}_{jd} + \vec{\alpha}_{jd} - \vec{\beta}_{j(d+1)}$. The GID posterior probability can be factorized as follows [71].

$$p\big(j \mid \mathcal{Y}_i, \vec{\pi}, \vec{\alpha}, \vec{\beta}\big) \propto \pi_j \prod_{l=1}^{D} p_{iBeta}\big(\mathcal{X}_{il} \mid \vec{\alpha_{jl}}, \vec{\beta_{jl}}\big) \tag{3.2}$$

where we have set $\mathcal{X}_{il} = \mathcal{Y}_{il}$ and $\mathcal{X}_{il} = \frac{\mathcal{Y}_{il}}{1+\sum_{k=1}^{l=1} \mathcal{Y}_{ik}}$ for $\vec{l} > 1$. $p_{iBeta}\big(\mathcal{X}_{il} \mid \vec{\alpha}_{jl}, \vec{\beta}_{jl}\big)$ is an inverted Beta distribution with parameters $\vec{\alpha}_{jl}$ and $\vec{\beta}_{jl}$ as below :

$$p_{iBeta}\big(\mathcal{X}_{il} \mid \vec{\alpha_{jl}}, \vec{\beta_{jl}}\big) = \frac{\Gamma\big(\alpha_{jd} + \beta_{jd}\big)}{\Gamma\big(\alpha_{jd}\big)\Gamma\big(\beta jd\big)} \mathcal{X}_{il}^{\alpha_{jl}-1} \big(1 + \mathcal{X}_{il}\big)^{-\big(\vec{\alpha}_{jl} + \vec{\beta}_{jl}\big)} \tag{3.3}$$

Let $\overrightarrow{\mathcal{Z}}_i = \big(\vec{Z_{i1}}, ..., \vec{Z_{iM}}\big)$ be a binary latent variable assigned to each observation $\overrightarrow{X}_i$. The values of $\mathcal{Z}_i$ satisfy $\mathcal{Z}_{ij} \in \{0, 1\}$, $\sum_{j=1}^{M} \mathcal{Z}_{ij} = 1$, $\mathcal{Z}_{ij} = 1$ if $\overrightarrow{X}_i$ belongs to component j and equal to 0, otherwise. The conditional distribution of latent variables $\mathcal{Z} = \big(\overrightarrow{Z}_1, ..., \overrightarrow{Z}_N\big)$ given the mixing coefficients $\overrightarrow{\pi}$, can be written as

$$p\big(\mathcal{Z} \mid \overrightarrow{\pi}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} \vec{\pi_j}^{Z_{ij}} \tag{3.4}$$

Thus, given the latent variables and the component parameters set we are able to write the conditional distribution of the data set $\mathcal{X} = \big(\overrightarrow{X}_1, ..., \overrightarrow{X}_N\big)$ as:

$$p\big(\mathcal{X} \mid \mathcal{Z}, \overrightarrow{\alpha}, \overrightarrow{\beta}\big) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\prod_{l=1}^{D} iBeta\big(\mathcal{X}_{il} \mid \vec{\alpha_{il}}, \vec{\beta_{jl}}\big) \right)^{Z_{ij}} \tag{3.5}$$

Feature selection is an important aspect when data is multidimensional and some features could be noisy, which can impact the algorithm performance as well as the clustering process. These features can thus be considered irrelevant since they do not have any discriminatory impact on the clustering. As so, to integrate feature selection with finite GID

mixture model in equation (3.2), and to take into consideration the fact that the features $\vec{X}_{il}$ are mostly not equally important for the clustering task, the following approximation for the $\mathcal{X}_{il}$ distribution has been suggested [71]:

$$p\big(\mathcal{X}_{il} \mid \mathcal{W}_{ikl}, \vec{\phi_{il}}, \vec{\alpha_{jl}}, \vec{\beta_{jl}}, \vec{\lambda_{kl}}, \vec{\tau_{kl}}\big) \simeq iBeta\big(\mathcal{X}_{il} \mid \vec{\alpha}_{jl}, \vec{\beta}_{jl})^{\phi_{il}}\big) \left(\prod_{K=1}^{K} iBeta\big(\mathcal{X}_{il} \mid \vec{\lambda_{kl}}, \vec{\tau_{kl}}\big)^{W_{ikl}} \right)^{1-\phi il} \tag{3.6}$$

where $\vec{\phi_{il}}$ is a binary latent variable, such that $\vec{\phi_{il}}$ = 1 indicates that l is relevant feature and follows an inverted Beta distribution iBeta $(\mathcal{X}_{il}|\alpha_{jl}, \beta_{jl})$. However, $\vec{\phi_{il}}$ = 0 represents that feature l is irrelevant and supposed to follow a finite mixture of inverted beta distributions independent from the class labels such as:

$$p\big(\mathcal{X}_{il}\big) = \sum_{K=1}^{K} \eta_{kl} iBeta\big(\mathcal{X}_{il} \mid \vec{\lambda_{kl}}, \vec{\tau_{kl}}\big) \tag{3.7}$$

where n_{kl} denotes a mixing probability and implies the prior probability that $\mathcal{X}_{il}$ is generated from the k^{th} component of the finite inverted beta mixture representing irrelevant features, and $\sum_{K=1}^{K} \eta_{kl}$ = 1.

In equation (3.6), $\mathcal{W}_{ikl}$ is a binary latent variable such that $\mathcal{W}_{ikl}$ = 1 only if $\mathcal{X}_{il}$ comes from the k^{th} component of the finite inverted beta mixture for the irrelevant features. The conditional distribution of the latent variables $\mathcal{W} = \big(\overrightarrow{W}_1, ..., \overrightarrow{W}_N\big)$ with $\overrightarrow{\mathcal{W}}_i = \big(\overrightarrow{W}_{i1}, ..., \overrightarrow{W}_{iK}\big)$ and $\overrightarrow{\mathcal{W}}_{ik} = \big(\overrightarrow{W}_{ik1}, ..., \overrightarrow{W}_{ikD}\big)$ given the mixing coefficients $\overrightarrow{\eta}$, can be written as

$$p\big(\mathcal{W} \mid \overrightarrow{\eta}\big) = \prod_{i=1}^{N} \prod_{K=1}^{K} \prod_{L=1}^{D} \vec{\eta_{kl}}^{W_{ikl}} \tag{3.8}$$

where $\overrightarrow{\eta} = \big(\overrightarrow{\eta}_1, ..., \overrightarrow{\eta}_K\big)$ with element $\overrightarrow{\eta}_k = \big(\overrightarrow{\eta}_{k1}, ..., \overrightarrow{\eta}_{kD}\big)$. The conditional distribution of the feature relevancy indicator variable $\overrightarrow{\phi} = \big(\overrightarrow{\phi}_1, ..., \overrightarrow{\phi}_N\big)$ with elements $\big(\overrightarrow{\phi}_{i1}, ..., \overrightarrow{\phi}_{iD}\big)$, given $\overrightarrow{\epsilon}$, is defined as

$$p\big(\overrightarrow{\phi} \mid \overrightarrow{\epsilon}\big) = \prod_{i=1}^{N} \prod_{l=1}^{D} \vec{\epsilon_{l_1}}^{\phi_{il}} \vec{\epsilon_{l_2}}^{1-\phi_{il}} \tag{3.9}$$

where ϕ is a Bernoulli variable such that $p(\phi_{il} = 1) = \epsilon_{l_1}$ and $p(\phi_{il} = 0) = \epsilon_{l_2}$. The vector $\overrightarrow{\epsilon} = \big(\overrightarrow{\epsilon}_1, ..., \overrightarrow{\epsilon}_D\big)$ represents the probabilities of the relevant features called feature saliencies such that $\overrightarrow{\epsilon}_l = (\epsilon_{l_1}, \epsilon_{l_2})$ and $(\epsilon_{l_1} + \epsilon_{l_2})$ = 1. Therefore, the likelihood of

the observed data set $\mathcal{X}$ following the finite GID mixture model with feature selection is given as follows :

$$p(\mathcal{X} \mid \mathcal{Z}, \mathcal{W}, \overrightarrow{\phi}, \overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\lambda}, \overrightarrow{\tau}) = \prod_{i=1}^{N}\prod_{j=1}^{M}\Bigg[\prod_{l=1}^{D} iBeta\big(\vec{X_{il}} \mid \vec{\alpha_{il}}, \vec{\beta_{jl}}\big)^{\phi_{il}} \times \Bigg(\prod_{K=1}^{K} iBeta\big(\vec{X_{il}} \mid \vec{\lambda_{kl}}, \vec{\tau_{kl}}\big)^{W_{ikl}}\Bigg]^{1-\phi_{il}}\Bigg)^{Z_{ij}} \tag{3.10}$$

The detailed description on this unsupervised feature selection model is given in [72].

3.1.1 Prior Specifications

The setting up of prior distributions is a very crucial step in variational learning. Hence, we have to place priors over ($\overrightarrow{\alpha}$), ($\overrightarrow{\beta}$) , ($\overrightarrow{\lambda}$) and ($\overrightarrow{\tau}$). The consideration of conjugate priors is the key factor which majorly simplifies variational inference method. In our case, we consider the gamma distribution to approximate a Beta distribution conjugate prior as suggested in [73] which gives the following priors:

$$p(\overrightarrow{\alpha}) = \mathcal{G}(\overrightarrow{\alpha} \mid \overrightarrow{u}, \overrightarrow{v}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \frac{\nu_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{3.11}$$

$$p(\overrightarrow{\beta}) = \mathcal{G}(\overrightarrow{\beta} \mid \overrightarrow{p}, \overrightarrow{q}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})} \beta_{jl}^{p_{jl}-1} e^{-q_{jl}\beta_{jl}} \tag{3.12}$$

$$p(\overrightarrow{\lambda}) = \mathcal{G}(\overrightarrow{\lambda} \mid \overrightarrow{g}, \overrightarrow{h}) = \prod_{K=1}^{M}\prod_{l=1}^{D} \frac{h_{kl}^{g_{kl}}}{\Gamma(g_{kl})} \lambda_{kl}^{g_{kl}-1} e^{-h_{kl}\lambda_{kl}} \tag{3.13}$$

$$p(\overrightarrow{\tau}) = \mathcal{G}(\overrightarrow{\tau} \mid \overrightarrow{s}, \overrightarrow{t}) = \prod_{K=1}^{M}\prod_{l=1}^{D} \frac{t_{sl}^{s_{kl}}}{\Gamma(s_{kl})} \tau_{kl}^{s_{kl}-1} e^{-t_{kl}\tau_{kl}} \tag{3.14}$$

where all the hyper-parameters $\overrightarrow{u}$ = $\{u_{jl}\}$, $\overrightarrow{v}$ = $\{v_{jl}\}$, $\overrightarrow{p}$ = $\{p_{jl}\}$, $\overrightarrow{q}$ = $\{q_{jl}\}$, $\overrightarrow{g}$= $\{g_{kl}\}$, $\overrightarrow{h}$= $\{h_{kl}\}$, $\overrightarrow{s}$ = $\{s_{kl}\}$ and $\overrightarrow{t}$ = $\{t_{kl}\}$ of the above conjugate priors are positive. We do not consider $\overrightarrow{\pi}$, $\overrightarrow{\eta}$ and $\overrightarrow{\epsilon}$ as random variables in our model so no priors are considered for them. The joint distribution of all the random variables for GID mixture model with feature selection is given by

$$p(\mathcal{X} \mid \mathcal{Z}, \mathcal{W}, \overrightarrow{\phi}, \overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\lambda}, \overrightarrow{\tau}) = p(\mathcal{X} \mid \mathcal{Z}, \mathcal{W}, \overrightarrow{\phi}, \overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\lambda}, \overrightarrow{\tau}) \times p(\mathcal{Z} \mid \overrightarrow{\pi}) p(\mathcal{W} \mid \overrightarrow{\eta}) p(\overrightarrow{\phi} \mid \overrightarrow{\epsilon}) p(\overrightarrow{\alpha}) p(\overrightarrow{\beta}) p(\overrightarrow{\lambda}) p(\overrightarrow{\tau}) \tag{3.15}$$

3.2 Online variational learning for finite generalized inverted Dirichlet mixture mode with feature selection

Variational procedures are very common and have been extensively utilized in the past to find approximations which are tractable for posterior distributions of a variety of statistical models [74]. One of the most integral part of designing finite mixture models is parameter estimation and to select the number of components correctly. In this section, we adopt an online variational framework of finite GID mixture model for parameter estimation and model selection. The online variational concept is taken into account for the dynamic nature of real-world data sets where the observations are sequential. Figure (3.1) represents the graphical representation for our model.

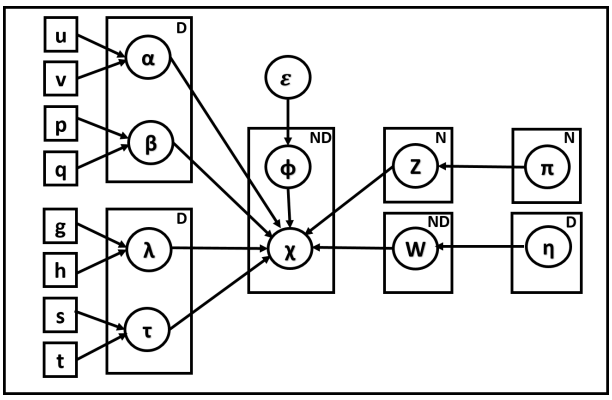

Figure 3.1: Graphical representation of finite GID mixture model with feature selection. The circles represent the random variables and model parameters. Numbers in the upper right corners of the plates indicate the number of repetitions.

The goal of variational inference method is to find a probability distribution $\mathcal{Q}(\Lambda)$ which approximates the true posterior distribution $p(\Lambda \mid \mathcal{X}, \gamma)$. We achieve this by maximizing the lower bound $\mathcal{L}$ on the evidence of model $p(\mathcal{X}|\gamma)$. This evidence of lower bound $\mathcal{L}$ is taken by applying Jensen's inequality on $p(\mathcal{X}|\gamma)$ [32] as:

$$\ln P(\mathcal{X} \mid \gamma) = \ln \int p(\mathcal{X} \mid \Lambda, \gamma) d\Lambda = \ln \int Q(\Lambda) \left(\frac{p(\mathcal{X} \mid \Lambda, \vec{\gamma})}{Q(\Lambda)} \right) d\Lambda$$
$$\geq \ln \int Q(\Lambda) \left(\frac{p(\mathcal{X} \mid \Lambda, \vec{\gamma})}{Q(\Lambda)} \right) d\Lambda = \mathcal{L}(\mathcal{Q}) \tag{3.16}$$

In theory, the lower bound $\mathcal{L}(\mathcal{Q})$ is maximized when $Q(\lambda) = p(\Lambda|\mathcal{X}, \Gamma)$. However, the actual posterior distribution is usually arithmetically intractable and cannot be directly utilized for variational inference. Hence, we use a factorization hypothesis to limit the form of $Q(\Lambda)$ in our work, such that $Q(\Lambda) = Q(\mathcal{Z})Q(\mathcal{W})Q(\overrightarrow{\phi})Q(\overrightarrow{\alpha})Q(\overrightarrow{\beta})Q(\overrightarrow{\lambda})Q(\overrightarrow{\tau})$. This hypothesis is commonly known as mean field approximation that comes out of statistical mechanics [75] and has been extensively utilized in the past for many applications (for example, [76]). It has been already described in the previous chapter.

The core idea is that as the model has conjugate priors, the functional form of the factors in the variational posterior distribution is known. According to this, by using general parametric forms on these distributions, the lower bound can be viewed as a function of the parameters of the distributions. We maximize the lower bound with respect to these parameters in order to obtain the optimization of variational factors. In our algorithm, the functional form of each factor is identical to its conjugate prior distribution, specifically discrete for $\mathcal{Z}$ and $\mathcal{W}$, bernoulli for $\overrightarrow{\phi}$, and gamma for $\overrightarrow{\alpha}$,$\overrightarrow{\beta}$, $\overrightarrow{\Lambda}$ and $\overrightarrow{\tau}$. Thus, the parametric forms of these variational posterior distributions could be defined as following:

$$Q(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}}, \quad Q(\mathcal{W}) = \prod_{i=1}^{N}\prod_{K=1}^{K}\prod_{L=1}^{D} \vec{m_{kl}}^{W_{ikl}} \tag{3.17}$$

$$Q(\overrightarrow{\phi}) = \prod_{j=1}^{N}\prod_{l=1}^{D} f_{il}^{\phi_{il}}(1-f_{il})^{1-\phi_{il}} \tag{3.18}$$

$$Q(\overrightarrow{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^{*}, \nu_{jl}^{*}), \quad Q(\overrightarrow{\beta}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\beta_{jl}|p_{jl}^{*}, q_{jl}^{*}) \tag{3.19}$$

$$Q(\overrightarrow{\lambda}) = \prod_{k=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\lambda_{kl}|g_{kl}^{*}, h_{kl}^{*}), \quad Q(\overrightarrow{\gamma}) = \prod_{k=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\gamma_{kl}|s_{kl}^{*}, t_{kl}^{*}) \tag{3.20}$$

We can obtain the parameterized lower bound $\mathcal{L}(\mathcal{Q})$ by substituting equations (3.17) - (3.20) into (3.16) as below :

$$\begin{aligned}\mathcal{L}(Q) &= \sum_{\theta}\int Q(\Theta,\Omega)\ln\left(\frac{p(\mathcal{X},\Theta,\Omega\mid\vec{\gamma})}{Q(\Theta,\Omega)}\right)d\Omega \\ &= \Big\langle \ln p(\mathcal{X},\Theta,\Omega\mid\vec{\gamma}) - \ln Q(\Theta,\Omega)\Big\rangle \end{aligned} \tag{3.21}$$

The detailed solution of the above equation is explained [32]. Then, the variational parameters r_{ij}, f_{il} and m_{ikl} can be calculated by maximizing $\mathcal{L}(\mathcal{Q})$ with respect to these parameters, respectively where,

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M} \tilde{r}_{ij}}, \quad f_{il} = \frac{\tilde{f}_{il}}{\tilde{f}_{il} + \hat{f}_{il}}, \quad m_{ikl} = \frac{\tilde{m}_{ikl}}{\sum_{k=1}^{k} \tilde{m}_{ikl}} \tag{3.22}$$

with

$$\tilde{r}_{ij} = exp\Bigg\{ \ln \pi_j + \sum_{l=1}^{D} \Bigg\{ f_{il}\big[\tilde{R}_{jl} + \big(\overline{\alpha}_{jl} - 1\big) \ln \mathcal{X}_{il} - (\overline{\alpha}_{jl} + \overline{\beta}_{jl}) \ln(1 + \mathcal{X}_{il})] \\ + (1 - f_{il}) \sum_{k=1}^{K} m_{ikl}\big[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl} - 1\big) \ln \mathcal{X}_{il} - (\overline{\lambda}_{kl} + \overline{\tau}_{kl}) \ln(1 + \mathcal{X}_{il})] \Bigg\} \Bigg\} \tag{3.23}$$

$$\tilde{m}_{ikl} = exp\Bigg\{ \ln \eta_{kl} + (1 - f_{il})\big[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl} - 1\big) \ln \mathcal{X}_{il} - (\overline{\lambda}_{kl} + \overline{\tau}_{kl}) \ln(1 + \mathcal{X}_{il})] \Bigg\} \tag{3.24}$$

$$\tilde{f}_{ij} = exp\Bigg\{ \ln \epsilon_{l_1} + \sum_{j=1}^{M} r_{ij}[\tilde{R}_{jl} + \big(\overline{\alpha}_{jl} - 1\big) \ln \mathcal{X}_{il} - (\overline{\alpha}_{jl} + \overline{\beta}_{jl}) \ln(1 + \mathcal{X}_{il})] \Bigg\} \tag{3.25}$$

$$\hat{f}_{il} = exp\Bigg\{ \ln \epsilon_{l_2} + \Bigg\{ \sum_{K=1}^{K} m_{ikl}[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl} - 1\big) \ln \mathcal{X}_{il} - (\overline{\lambda}_{k1} + \overline{\tau}_{k1}) \ln(1 + \mathcal{X}_{il})] \Bigg\} \Bigg\} \tag{3.26}$$

$$\begin{aligned} \tilde{R} = \ln & \frac{\Gamma\big(\tilde{\alpha} + \tilde{\beta}\big)}{\Gamma\big(\overline{\beta}\big)\Gamma\big(\overline{\alpha}\big)} \\ & + \tilde{\alpha}\Big[\psi\big(\overline{\alpha} + \tilde{\beta}\big) - \psi\big(\overline{\alpha}\big)\Big]\Big[\big\langle \ln \alpha \big\rangle - \ln \overline{\alpha}\Big] \\ & + \tilde{\beta}\Big[\psi\big(\overline{\beta} + \tilde{\alpha}\big) - \psi\big(\overline{\beta}\big)\Big]\Big[\big\langle \ln \beta \big\rangle - \ln \overline{\beta}\Big] \\ & + 0.5\tilde{\alpha}^2\Big[\psi\big(\overline{\alpha} + \tilde{\beta}\big) - \psi\big(\overline{\alpha}\big)\Big]\Big[\big\langle \ln \alpha \big\rangle - \ln \overline{\alpha}\Big]^2 \\ & + 0.5\tilde{\beta}^2\Big[\psi\big(\overline{\beta} + \tilde{\alpha}\big) - \psi\big(\overline{\beta}\big)\Big]\Big[\big\langle \ln \beta \big\rangle - \ln \overline{\beta}\Big]^2 \\ & + \tilde{\alpha}\tilde{\beta}\psi\big(\overline{\alpha} + \tilde{\beta}\big)\Big[\big\langle \ln \beta \big\rangle - \ln \overline{\beta}\Big]\Big[\big\langle \ln \alpha \big\rangle - \ln \overline{\alpha}\Big] \end{aligned} \tag{3.27}$$

$$
\begin{aligned}
\tilde{F} = \ln \frac{\Gamma\big(\tilde{\lambda}+\tilde{\tau}\big)}{\Gamma\big(\overline{\tau}\big)\Gamma\big(\overline{\lambda}\big)} \\
&+ \tilde{\lambda}\Big[\psi\big(\overline{\lambda}+\tilde{\tau}\big) - \psi\big(\overline{\lambda}\big)\Big]\Big[\big\langle \ln\lambda\big\rangle - \ln\overline{\lambda}\Big] \\
&+ \tilde{\tau}\Big[\psi\big(\overline{\tau}+\tilde{\lambda}\big) - \psi\big(\overline{\tau}\big)\Big]\Big[\big\langle \ln\tau\big\rangle - \ln\overline{\tau}\Big] \\
&+ 0.5\tilde{\lambda}^2\Big[\psi'\big(\overline{\lambda}+\tilde{\tau}\big) - \psi'\big(\overline{\lambda}\big)\Big]\Big[\big\langle \ln\lambda\big\rangle - \ln\overline{\lambda}\Big]^2 \\
&+ 0.5\tilde{\tau}^2\Big[\psi'\big(\overline{\tau}+\tilde{\lambda}\big) - \psi'\big(\overline{\tau}\big)\Big]\Big[\big\langle \ln\tau\big\rangle - \ln\overline{\tau}\Big]^2 \\
&+ \tilde{\lambda}\tilde{\tau}\psi'\big(\overline{\lambda}+\tilde{\tau}\big)\Big[\big\langle \ln\lambda\big\rangle - \ln\overline{\lambda}\Big]\Big[\big\langle \ln\tau\big\rangle - \ln\overline{\tau}\Big]
\end{aligned}
\tag{3.28}
$$

where $\psi(.)$ is the digamma function that is defined as $\psi(\alpha) = \dfrac{d\ln\Gamma(\alpha)}{d(\alpha)}$.

Similarly, we can obtain the update equations of the hyper-parameters of variational factors α, β, γ and τ. Finally, the mixing coefficients π_{ij}, η_{kl} and the feature salencies ϵ_{l_1} can be calculated as :

$$
\pi_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij}, \quad \eta_{kl} = \frac{1}{N}\sum_{i=1}^{N} m_{ikl}, \quad \epsilon_{l_1} = \frac{1}{N}\sum_{i=1}^{N} f_{il} \tag{3.29}
$$

In this subsection, we propose an online variational learning framework with unsupervised feature selection for finite GID mixture model for sequential data . The proposed algorithm approach of online learning is based upon the variational technique developed in [77] which we consider in our work. The approach has been already described in the previous section.

The core idea of the online variational algorithm is to maximize the present variational lower bound successively. Suppose that we have already observed the data set $\mathcal{X}^{(\iota-1)}$ = $(X_1, \ldots, X_{(\iota-1)})$ and determined the variational factors $\mathcal{Q}(\overrightarrow{\phi}_{(\iota-1)})$, $\mathcal{Q}(\overrightarrow{Z}_{(\iota-1)})$, $\mathcal{Q}(\overrightarrow{\mathcal{W}}_{(\iota-1)})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\alpha})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\beta})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\lambda})$ and $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\tau})$ as well as the parameters $\overrightarrow{\pi}^{(\iota-1)}$, $\overrightarrow{\eta}^{(\iota-1)}$, $\overrightarrow{\epsilon}^{(\iota-1)}$. When the newly arriving data $X\iota$ is observed, we need to update the current ι^{th} optimal value for a variational factor according to the $((\iota-1)^{th})$ values of the other variational factors. Later, we update the ι^{th} optimal value for the second factor by holding the newly obtained ι^{th} value of the first factor fixed and setting other factors still to their $(\iota-1)^{th}$ values. We keep repeating this procedure until all the variational factors are

updated with respect to the new observation .

In this work, we first maximize the current lower bound $\mathcal{L}^{(t)}\mathcal{Q}$ with respect to $\mathcal{Q}(\overrightarrow{\phi}_{\iota})$, while other variational factors are set to $\mathcal{Q}(\overrightarrow{Z}_{(\iota-1)})$, $\mathcal{Q}(\overrightarrow{\mathcal{W}}_{(\iota-1)})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\alpha})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\beta})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\lambda})$ and $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\tau})$, and the feature saliency $\overrightarrow{\epsilon}$ is set to $\overrightarrow{\epsilon}^{(\iota-1)}$. Hence, the variational solution for $Q(\overrightarrow{\phi}_{\iota})$ can be calculated as

$$Q(\overrightarrow{\phi}) = \prod_{l=1}^{D} f_{\iota l}^{\phi_{\iota l}}(1-f_{\iota l})^{1-\phi_{\iota l}} \tag{3.30}$$

where:

$$f_{\iota l} = \frac{\tilde{f}_{\iota l}}{\tilde{f}_{\iota l}+\hat{f}_{\iota l}}, \tag{3.31}$$

In the above equation (3.31), we substitute the below values of (3.32) and (3.33) by modifying the equations (3.25) and (3.26) respectively to

$$\tilde{f}_{\iota l} = exp\Bigg\{ \ln \epsilon_{l_1}^{(\iota-1)} + \sum_{j=1}^{M} r_{(\iota-1)j}[\tilde{R}_{jl} + \big(\overline{\alpha}_{jl}-1\big)\ln \mathcal{X}_{\iota l} - (\overline{\alpha}_{jl}+\overline{\beta}_{jl})\ln(1+\mathcal{X}_{\iota l})] \Bigg\} \tag{3.32}$$

$$\hat{f}_{\iota l} = exp\Bigg\{ \ln \epsilon_{l_2}^{(\iota-1)} + \Bigg\{ \sum_{K=1}^{K} m_{(\iota-1)kl}[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl}-1\big)\ln \mathcal{X}_{\iota l} - (\overline{\lambda}_{k1}+\overline{\tau}_{k1})\ln(1+\mathcal{X}_{\iota l})] \Bigg\} \Bigg\} \tag{3.33}$$

In the next step, we maximize the current lower bound $\mathcal{L}^{(t)}\mathcal{Q}$ with respect to $\mathcal{Q}(\overrightarrow{\mathcal{Z}}_{\iota})$, while $\mathcal{Q}(\overrightarrow{\phi}_{\iota})$ is fixed, $\overrightarrow{\pi}$ is set to $\overrightarrow{\pi}^{(\iota-1)}$, $\mathcal{Q}(\overrightarrow{\alpha})$, $\mathcal{Q}(\overrightarrow{\beta})$, are set to $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\alpha})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\beta})$, respectively. Based on equation (3.17), the variational solution for $\mathcal{Q}(\overrightarrow{\mathcal{Z}}_{\iota})$ is given by

$$\mathcal{Q}\big(\overrightarrow{\mathcal{Z}}_{\iota}\big) = \prod_{j=1}^{M} \tilde{r}_{\iota j}^{Z_{\iota j}} \tag{3.34}$$

where,

$$r_{\iota j} = \frac{\tilde{r}_{\iota j}}{\sum_{j=1}^{M}\tilde{r}_{\iota j}}, \tag{3.35}$$

We modify the equation (3.23) discussed in the previous section to the one below for online variational case

$$\tilde{r}_{\iota j} = exp\Bigg\{ \ln \pi_j{}^{(\iota-1)} + \sum_{l=1}^{D} \Bigg\{ f_{\iota l}\big[\tilde{R}_{jl} + \big(\overline{\alpha}_{jl} - 1\big) \ln \mathcal{X}_{\iota l} - (\overline{\alpha}_{jl} + \overline{\beta}_{jl}) \ln(1 + \mathcal{X}_{\iota l})] \\ +(1 - f_{il}) \sum_{k=1}^{K} m_{(\iota-1)kl}\big[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl} - 1\big) \ln \mathcal{X}_{\iota l} - (\overline{\lambda}_{kl} + \overline{\tau}_{kl}) \ln(1 + \mathcal{X}_{\iota l})] \Bigg\} \Bigg\} \tag{3.36}$$

Subsequently, we maximize $\mathcal{L}^{(t)}\mathcal{Q}$ with respect to $\mathcal{Q}(\overrightarrow{\mathcal{W}}_\iota)$, using $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\lambda})$, $\mathcal{Q}^{(\iota-1)}(\overrightarrow{\tau})$ and $\overrightarrow{\eta}^{(\iota-1)}$, while $\mathcal{Q}(\overrightarrow{\phi}_\iota)$ is considered fixed, such that

$$\mathcal{Q}\big(\overrightarrow{\mathcal{W}_\iota}\big) = \prod_{K=1}^{K} \prod_{L=1}^{D} \vec{m_{\iota kl}}^{W_{\iota kl}} \tag{3.37}$$

where,

$$m_{\iota kl} = \frac{\tilde{m}_{\iota kl}}{\sum_{k=1}^{k} \tilde{m}_{\iota kl}} \tag{3.38}$$

The equation (3.24) in the previous section is modified as below

$$\tilde{m}_{\iota kl} = exp\Bigg\{ \ln \eta_{kl}^{(\iota-1)} + (1 - f_{(ul)})\big[\tilde{F}_{kl} + \big(\overline{\lambda}_{kl} - 1\big) \ln X_{\iota l} - (\overline{\lambda}_{kl} + \overline{\tau}_{kl}) \ln(1 + X_{\iota l})] \Bigg\} \tag{3.39}$$

Now in order to obtain the variational solution for $\mathcal{Q}^{(\iota)}(\overrightarrow{\alpha})$, we need to maximize $\mathcal{L}^{(t)}\mathcal{Q}$ with respect to the variational factor $\mathcal{Q}^{(\iota)}(\overrightarrow{\alpha})$, while holding $\mathcal{Q}(\overrightarrow{\phi}_\iota)$ and $\mathcal{Q}(\overrightarrow{\mathcal{Z}}_\iota)$ fixed as

$$\mathcal{Q}^{\iota}(\overrightarrow{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}^{(\iota)} | u_{jl}^{*(\iota)}, \nu_{jl}^{*(\iota)}) \tag{3.40}$$

A significant characteristic of the adopted variational method [52], which cites that variational inference could be handled as a normal gradient method [56] whixh has been described in the previous chapter. In this case, the natural gradients of the variational hyper-parameters u_{jl}^{*} and v_{jl}^{*} are structurally equivalent to the updates given by

$$\Delta u_{jl}^{*(\iota)} = u_{jl}^{*(\iota)} - u_{jl}^{*(\iota-1)} = u_{j\iota} + Nr_{\iota j} f_{\iota j} \overline{\alpha}_{jl} \big[\psi'\big(\overline{\alpha}_{\iota j} + \tilde{\beta}_{\iota j}\big) - \psi\big(\overline{\alpha}_{jl}\big)\big] + \overline{\beta}_{jl} \big[\psi'\big(\overline{\alpha}_{\iota j} + \tilde{\beta}_{\iota j}\big) \Big[\big\langle \ln \beta_{jl} \big\rangle - \ln \overline{\beta}_{jl}\Big] - u_{jl}^{*(\iota-1)} \tag{3.41}$$

$$\Delta \nu_{jl}^{*(\iota)} = \nu_{jl}^{*(\iota)} - \nu_{jl}^{*(\iota-1)} = \nu_{j\iota} - Nr_{\iota j} f_{\iota l} \ln \frac{\mathcal{X}_{\iota t}}{1 + \mathcal{X}_{\iota t}} - \nu_{jl}^{*(\iota-1)} \tag{3.42}$$

Thus, the variational solutions to hyper parameters $u_{jl}^{*(\iota)}$ and $v_{jl}^{*(\iota)}$ are calculated through their natural gradients as

$$u_{jl}^{*(\iota)} = u_{jl}^{*(\iota-1)} + \rho_\iota \Delta u_{jl}^{*(\iota)} \tag{3.43}$$

$$\nu_{jl}^{*(\iota)} = \nu_{jl}^{*(\iota-1)} + \rho_\iota \Delta \nu_{jl}^{*(\iota)} \tag{3.44}$$

where ρ_ι is the learning rate and is defined as

$$\rho_\iota = (\delta_o + \iota)^{-\epsilon} \tag{3.45}$$

with the constraints:$\xi \epsilon (0.5, 1]$and $\delta_o \geq 0$. The function of the learning rate here is adopted from [57], described in detail in the previous chapter. Similarly, the variational factors $\mathcal{Q}^{(\iota)}(\overrightarrow{\beta})$, $\mathcal{Q}^{(\iota)}(\overrightarrow{\lambda})$, $\mathcal{Q}^{(\iota)}(\overrightarrow{\tau})$, are updated as

$$\mathcal{Q}^{(\iota)}\overrightarrow{(\beta)} = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\beta_{jl}^{(\iota)} | p_{jl}^{*(\iota)}, q_{jl}^{*(\iota)}) \tag{3.46}$$

$$\mathcal{Q}^{(\iota)}\overrightarrow{(\lambda)} = \prod_{k=1}^{K} \prod_{l=1}^{D} \mathcal{G}(\lambda_{kl}^{(\iota)} | g_{kl}^{*(\iota)}, h_{kl}^{*(\iota)}) \tag{3.47}$$

$$\mathcal{Q}^{(\iota)}\overrightarrow{(\tau)} = \prod_{k=1}^{K} \prod_{l=1}^{D} \mathcal{G}(\tau_{kl}^{(\iota)} | s_{kl}^{*(\iota)}, t_{kl}^{*(\iota)}) \tag{3.48}$$

where

$$p_{jl}^{*(\iota)} = p_{jl}^{*(\iota-1)} + \rho_\iota \Delta p_{jl}^{*(\iota)}, \quad q_{jl}^{*(\iota)} = q_{jl}^{*(\iota-1)} + \rho_\iota \Delta q_{jl}^{*(\iota)} \tag{3.49}$$

$$g_{kl}^{*(\iota)} = g_{kl}^{*(\iota-1)} + \rho_\iota \Delta g_{kl}^{*(\iota)}, \quad h_{kl}^{*(\iota)} = h_{kl}^{*(\iota-1)} + \rho_\iota \Delta h_{kl}^{*(\iota)} \tag{3.50}$$

$$s_{kl}^{*(\iota)} = s_{kl}^{*(\iota-1)} + \rho_\iota \Delta s_{kl}^{*(\iota)}, \quad t_{kl}^{*(\iota)} = t_{kl}^{*(\iota-1)} + \rho_\iota \Delta t_{kl}^{*(\iota)} \tag{3.51}$$

The corresponding natural gradients of the variational hyper parameters in the above equations are given by

$$\begin{aligned}\Delta p_{jl}^{*(\iota)} = p_{jl}^{*(\iota)} - p_{jl}^{*(\iota-1)} = p_{j\iota} + N r_{\iota j} f_{\iota l} \overline{\beta}_{jl} \big[\psi\big(\overline{\alpha}_{\iota j} + \overline{\beta}_{\iota j}\big) - \psi\big(\overline{\beta}_{jl}\big)\big] \\ + \overline{\alpha}_{jl} \big[\psi'\big(\overline{\alpha}_{\iota j} + \overline{\beta}_{\iota j}\big)\big] \Big[\big\langle \ln \alpha_{jl} \big\rangle - \ln \overline{\alpha}_{jl}\Big] - p_{jl}^{*(\iota-1)}\end{aligned} \tag{3.52}$$

$$\Delta q_{jl}^{*(\iota)} = q_{jl}^{*(\iota)} - q_{jl}^{*(\iota-1)} = q_{j\iota} + N r_{\iota j} f_{\iota l} \ln \frac{1}{1 + \mathcal{X}_{\iota l}} - q_{jl}^{*(\iota-1)} \tag{3.53}$$

$$\begin{aligned}\Delta g_{kl}^{*(\iota)} = g_{kl}^{*(\iota)} - g_{kl}^{*(\iota-1)} = g_{kl} + N(1 - \phi_{\iota l}) m_{\iota k l} \overline{\lambda}_{kl} \big[\psi\big(\overline{\lambda}_{kl} + \overline{\tau}_{kl}\big) \\ - \psi\big(\overline{\lambda}_{kl}\big) + \overline{\tau}_{kl} \big[\psi'\big(\overline{\lambda}_{kl} + \overline{\tau}_{kl}\big) \Big[\big\langle \ln \tau_{kl} \big\rangle - \ln \overline{\tau}_{kl}\Big] - g_{jl}^{*(\iota-1)}\end{aligned} \tag{3.54}$$

$$\Delta h_{kl}^{*(\iota)} = h_{kl}^{*(\iota)} - h_{kl}^{*(\iota-1)} = h_{kl} - N(1 - \phi_{\iota l}) m_{\iota k l} \ln \frac{\mathcal{X}_{\iota l}}{1 + \mathcal{X}_{\iota t}} - h_{kl}^{*(\iota-1)} \tag{3.55}$$

$$\begin{aligned}\Delta s_{kl}^{*(\iota)} = s_{kl}^{*(\iota)} - s_{kl}^{*(\iota-1)} = s_{kl} + N(1 - \phi_{\iota l}) m_{\iota k l} \overline{\tau}_{kl} \big[\psi\big(\overline{\lambda}_{kl} + \overline{\tau}_{kl}\big) \\ - \psi\big(\overline{\tau}_{kl}\big) + \overline{\lambda}_{kl\psi'} \big(\overline{\lambda}_{kl} + \overline{\tau}_{kl}\big) \big(\big\langle \ln \lambda_{kl} \big\rangle - \ln \overline{\lambda}_{kl}\big)\big] - s_{jl}^{*(\iota-1)}\end{aligned} \tag{3.56}$$

$$\Delta t_{kl}^{*(\iota)} = t_{kl}^{*(\iota)} - t_{kl}^{*(\iota-1)} = t_{kl} - N(1 - \phi_{kl})) m_{kl} \ln \frac{1}{1 + \mathcal{X}_{\iota t}} - t_{kl}^{*(\iota-1)} \tag{3.57}$$

Finally, we can update variational parameters $\overrightarrow{\pi}^{(\iota)}$, $\overrightarrow{\eta}^{(\iota)}$ and $\overrightarrow{\epsilon}^{(\iota)}$ as

$$\pi_j^{(\iota)} = \pi_j^{(\iota-1)} + \rho_\iota \Delta \pi_j \tag{3.58}$$

$$\eta_{kl}^{(\iota)} = \eta_{kl}^{(\iota-1)} + \rho_\iota \Delta\eta_{kl} \tag{3.59}$$

$$\epsilon_{l_1}^{(\iota)} = \epsilon_{l_1}^{(\iota-1)} + \rho_\iota \Delta\epsilon_{l_1}^{\iota} \tag{3.60}$$

where the natural gradients $\Delta\pi_j^{(\iota)}$, $\Delta\eta_{kl}^{(\iota)}$ and $\Delta\epsilon_{l_1}^{(\iota)}$ are calculated by

$$\Delta\pi_j^{(\iota)} = \pi_j^{(\iota)} - \pi_j^{(\iota-1)} = \Big(\frac{N}{t}\Big) r_{\iota j} - \pi_j^{(\iota-1)} \tag{3.61}$$

$$\Delta\eta_{kl}^{(\iota)} = \eta_{kl}^{(\iota)} - \eta_{kl}^{(\iota-1)} = \Big(\frac{N}{t}\Big) m_{\iota kl} - \eta_{kl}^{(\iota-1)} \tag{3.62}$$

$$\Delta\epsilon_{l_1}^{(\iota)} = \epsilon_{l_1}^{(\iota)} - \epsilon_{l_1}^{(\iota-1)} = \Big(\frac{N}{t}\Big) f_{\iota l} - \epsilon_{l_1}^{(\iota-1)} \tag{3.63}$$

Furthermore, as showed in [77], the online variational algorithm can be defined as a stochastic approximation method [78] in order to estimate the expected lower bound and the convergence is assured if the learning standard satisfies these conditions:

$$\sum_{\iota=1}^{\infty} \rho_\iota = \infty, \sum_{i=1}^{\infty} \rho_\iota^2 < \infty \tag{3.64}$$

The major cause of slow convergence is the affect on later estimations due to inaccurate hyper parameter estimations which occur in the earlier inference stages. Therefore, including the learning rate in the learning process is considered important for accelerating the convergence rate. The steps for online variational inference for finite GID mixture model with feature selection are abstracted in algorithm [3.2].

Algorithm 1 Online Variational learning of the finite GID mixture model with feature selection

---

1. Choose the initial number of components M and K.
2. Initialize the values for hyper-parameters u_{jl}, ν_{jl}, p_{jl}, q_{jl}, g_{kl}, h_{kl}, s_{kl}, t_{kl}.
3. Using K-means algorithm, initialize the values of r_{ij} and m_{ikl}.
4. **for** $t = 1 \rightarrow N$ **do**

    i The variational E-step:

ii Update the variational solutions for $\mathcal{Q}(\overrightarrow{\phi}_\iota)$, $\mathcal{Q}(\overrightarrow{\mathcal{Z}}_\iota)$ and $\mathcal{Q}(\overrightarrow{\mathcal{W}}_\iota)$ through equations (3.30), (3.34) and (3.37) respectively.

iii The variational M-step:

iv Compute learning rate $\rho_\iota = (\delta_o + \iota)^{-\epsilon}$ as in equation (3.45)

v Calculate the natural gradients $\Delta u_{jl}^{*(\iota)}$, $\Delta \nu_{jl}^{*(\iota)}$, $\Delta p_{jl}^{*(\iota)}$, $\Delta q_{jl}^{*(\iota)}$, $\Delta g_{kl}^{*(\iota)}$, $\Delta h_{kl}^{*(\iota)}$, $\Delta s_{kl}^{*(\iota)}$ and $\Delta t_{kl}^{*(\iota)}$ using equations (3.41),(3.42) , (3.52), (3.53),(3.54), (3.55), (3.56),(3.57) respectively

vi Update the variational solution for $\mathcal{Q}^\iota(\overrightarrow{\alpha})$, $\mathcal{Q}^{(\iota)}(\overrightarrow{\beta})$, $\mathcal{Q}^{(\iota)}(\overrightarrow{\lambda})$ $\mathcal{Q}^{(\iota)}(\overrightarrow{\tau})$ through equations (3.40) ,(3.46) , (3.47), (3.48) and (3.57)

vii Calculate the natural gradients $\Delta\pi_j^{(\iota)}$, $\Delta\eta_{kl}^{(\iota)}$ and $\Delta\epsilon_{l_1}^{(\iota)}$ via equations (3.61), (3.62), (3.63) respectively, for parameters $\overrightarrow{\pi}^{(\iota)}$, $\overrightarrow{\eta}^{(\iota)}$ and $\overrightarrow{\epsilon}^{(\iota)}$

viii Update the current solutions for $\overrightarrow{\pi}^{(\iota)}$, $\overrightarrow{\eta}^{(\iota)}$ and $\overrightarrow{\epsilon}^{(\iota)}$ using equations (3.58), (3.59), (3.60)

ix Repeat the variational E-step and M-step until new data is observed.

5. **end for**

3.3 Experimental results

In this section, we investigate the efficiency of our proposed online variational GID mixture model with feature selection by synthetic data and three challenging medical applications. The synthetic data purpose is to examine the online variational algorithm accuracy in terms of estimation of parameters and model selection. We performed medical image segmentation and feature selection on three data sets of different diseases and different medical image testing techniques. We applied the algorithm to detect brain tumor, skin lesion and computer aid detection (CAD) of malaria. Furthermore, we have used two different formats of images to test the applicability of the algorithm on varied output formats, namely, MRI scans, dermoscopic photographs. The main goal to focus on medical applications was to visualize the way different analytical and statistical mixture model methods can help the healthcare industry to give more precise results while diagnosing any patients health using machine learning.

Concerning the medical data sets we have used for the experiments, we make a performance comparison of our algorithm of online variational learning of finite GID with feature selection (OVGIDMM) with two other models namely online variational learning of finite inverted dirichlet mixture model (OVIDMM) and OVGMM to illustrate the merits of our algorithm implementation. OVIDMM is considered for comparison since it has less co-variance compared to our proposed algorithm and can also be used for positive vectors . OVGMM is considered since it is an extremely popular and novel approach. The below sections would follow the description on image segmentation, feature selection and the results obtained by calculating different evaluation metrics.

3.3.1 Image segmentation

Image segmentation is considered as an integral part for computer vision. It is the process used to partition the image into many segments according to the pixels. The main aim of the segmentation process is to change the representation of the image to make the analysis and interpretation process easier, since we get more understanding about the image and to detect the lines or curves in the image. In other words, the image segmentation makes a label for each pixel in order to have a table of similar features. Each pixel is similar to the other computed features like color or texture.

There are mainly two types of image segmentation techniques called non - contextual thresholding and contextual thresholding. The non-contextual type doesn't consider the spatial relationships between features in the image but the contextual technique does consider these relationships for example grouping together pixels with similar grey levels. In all our experiments in this chapter, we used the non contextual thresholding technique called RGB colour thresholding. The input to the thresholding operation was typically gray scale for brain tumor detection and color scale for the skin melanoma and CAD of malaria. In this implementation, the output is a binary image representing the segmentation where the black pixels correspond to background and white pixels correspond to foreground (or vice versa). The detection of edges in various clusters formed by the image segmentation helped us to derive the diagnostic insights to it by comparing it with the ground truth. The major challenge while performing segmentation was to identify the pixels that belong to features of interest to us. As an example, we performed the followings steps to detect the brain tumor by MRI image segmentation :- where we first extracted the brain structures and

then did localization of tumor region of interest (ROI) and then considering the size of the tumor with other structures in the brain and then diagnosed the tumor by comparing it to the ground truth.

In automated MRI image analysis, image segmentation is considered to be a preliminary step. The different types of factors which can affect on deciding the segmentation type when dealing with medical data sets are; which main body part is being considered, the imaging technique and lastly the application type for deciding the best suitable segmentation [79]. The applications in the healthcare field could be related to cell counting, measurement process for organ, counting of cells or prediction of abnormal growth which would depend on boundary extraction. There are a few general challenges that could be experienced when dealing with medical image segmentation: 1) the variability in sensing of the main part is very large, especially because it is very complicated when dealing with human anatomy, 2) the affect of medical image is different for each organ of the body since the motion of the heart also affects the imaging quality, 3) the noise effect of the sensor being used for detection.

In our model we extracted the feature of each image using the most commonly used technique of color histogram where we calculated the green color component histogram value for an RGB component of an image since the red and blue colour component had no varitions and followed no statistical model. Color is one of the most outstanding features of the image, it is the most important human visual content and it is very easy to calculate. The color histogram for an image is constructed by quantizing the colors within the image and counting the number of pixels of each color. Then, we take a summation of it and find the mean and standard deviation from the color histogram. Finally, it is stored in a 1D array. This value is calculated for every image in the data set [80].

3.3.2 Synthetic data

Our proposed algorithm was evaluated by quantitative analysis on dimensional data with two relevant features. These data sets have different data sizes namely, 200, 600, 900 and 1200. The relevant features were created in the converted space from mixture of inverted beta distributions with well-separated components. The table [3.1] below demonstrates the actual and evaluated parameters of the distributions using our proposed online variational

approach and considering the relevant features for each data set. According to the results obtained, the model parameters representing relevant features, and its mixing coefficients are precisely estimated by our online algorithm. In our experiments for synthetic data , the components M and K number had been initialized with 6 and 2 for two dimensional data respectively with equivalent mixing coefficients and the feature salencies value are initialized at 0.5.The initial values of the hyper-parameters u, p, g and s for the conjugate priors are fixed to 1, v to 0.04, q to 0.03, h to 0.05 and t to 0.06. ϵ and Σ the learning rate parameters are fixed to 0.5 and 64.

Data set	N_j	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	π_j	$\hat{\alpha_{j1}}$	$\hat{\beta_{j1}}$	$\hat{\alpha_{j2}}$	$\hat{\beta_{j2}}$	$\hat{\pi}_j$
Data set 1	100	1	20	13	18	15	0.5	19.91	13.05	18.99	15.30	0.50
(N = 200)	100	2	25	15	22	12	0.5	25.43	15.27	21.76	12.27	0.50
Data set 2	200	1	20	13	24	15	0.33	21.59	14.53	23.74	15.13	0.33
(N=600)	200	2	22	15	25	12	0.33	21.19	15.11	25.41	12.76	0.33
	200	3	25	16	22	14	0.34	24.79	16.10	23.86	13.86	0.34
Data set 3	300	1	20	13	24	15	0.33	20.68	13.51	24.17	14.13	0.33
(N=900)	300	2	22	15	25	12	0.33	21.89	14.77	24.46	13.26	0.33
	300	3	21	15	22	14	0.34	20.53	15.05	22.82	13.89	0.34
Data set 4	400	1	20	13	20	15	0.33	20.06	14.37	21.49	13.88	0.33
(N=1200)	400	2	22	15	20	12	0.33	21.72	14.96	20.83	14.21	0.33
	400	3	21	15	22	14	0.34	20.89	13.89	23.02	14.28	0.34

Table 3.1: Real and estimated parameters of different data sets. N denotes the total number of data points, N_j denotes the number of data points in the cluster j. $\alpha_{j1}, \beta_{j1}, \alpha_{j2}, \beta_{j2}$ and π_j are the real parameters and $\hat{\alpha}_{j1}, \hat{\beta}_{j1}, \hat{\alpha}_{j2}, \hat{\beta}_{j2},$ and $\hat{\pi}_j$ are the parameters estimated by our proposed algorithm.

3.3.3 Medical image data sets

After validating the algorithm on synthetic data sets, we applied it on three challenging medical data sets for brain tumor detection, skin melanoma detection and CAD of malaria data set. We observed that our algorithm could detect the morphological and structural anomalies similar to the ground truth data when performing image segmentation. We used

30 different patient images in each case of image segmentation for brain tumor detection and skin melanoma and compared the result of our proposed algorithm OVGIDMM with OVIDMM and OVGMM by taking out the mean of all the images values obtained in terms of Adjusted Rand Index (ARI) score, Adjusted Mutual Information (AMI) score, V-Measure score, Fowlkes - Mallows (FM) index, Dice similarity coefficient and Jaccard similarity index for evalution of the accuracy. The evaluation of CAD of malaria data set was also done by comparing our algorithm to OVIDMM and OVGMM on the basis of confusion matrix for classification of the patients into uninfected and parasitized category.

In our experiments for image segmentation of brain MRI images section [3.3.3] and skin melanoma images section [3.3.3], the number of components M and K had been initialized with 16 and 2 and the feature salencies value were initialized at 0.5. The initial values of the hyper-parameters u, p, g and s for the conjugate priors are fixed to 1, v to 0.03, q to 0.035, s to 0.05 and t to 0.06. ϵ and Σ the learning rate parameters are fixed to 0.5 and 64. The initialization was kept different for the testing of CAD of malaria data set which has been described in section [3.3.3].

Brain tumor detection

Tumor results from any abnormal proliferation of different kinds of cells in the body and can be either benign or malignant [81]. Brain tumor is accounted by the occurrence of the tumor in the brain or the skull. Benign brain tumor has uniformity in structure and does not contain proliferative cells, while malignant brain tumors have non-uniform (heterogeneous) structure and contain proliferative cells. Further to this, brain tumor is divided into two categories: primary and secondary. Primary tumors begin in the brain tissue while secondary spread from other tissues to the brain. According to the World Health Organization and the American Brain Tumor Association, the tumor types are classified on the scale of grade I to IV representing benign and malignant tumors. Benign tumors are grade I and grade II gliomas while the malignant are grade III and IV gliomas. Grade I and II gliomas are also called low grade tumor type and have slow growth, while grades III and IV are called high grade tumor type and have fast tumor growth. Gliomas and meningiomas are examples of low-grade primary tumors and are classified as benign tumors. Glioblastoma and astrocytomas on the other hand, are a class of high-grade primary tumors and are therefore classified as malignant tumors [82].

High grade gliomas (HGG) are the most frequently diagnosed primary brain tumor. Despite decades of research, HGG are among the top 10 causes of cancer deaths. The prognosis is quick and the life expectancy of a patient diagnosed with glioblastoma is drastically reduced. An estimate of 13, 000 people have been reported to die due to brain tumors [83]. Patients with gliomas are kept under serial monitoring and magnetic resonance imaging (MRI) or computed tomography (CT) observations on the tumor growth are made every 6 to 12 months. Brain cancer can affect any individual at any age and its impact on the body may not be the same for every individual [84].

MRI is routinely employed as a non invasive imaging method to characterize brain tumors and give pretreatment evaluations on it [85]. This image can be further segmented where the process of segmentation involves identifying and separating tumor micro-environment tissues, such as edema and dead cells, from normal brain tissues [86]. Several researchers have proposed various methodologies and algorithms for brain tumor segmentation by using K - means clustering technique [87], Spatial Fuzzy C-means [88], convolution neural network (CNN) as pixel classifier for the segmentation process [89] and K-Medoids clustering [90].

In this chapter, the brain tumor data set was obtained from kaggle [1]. The data set consisted of 110 brain MRI images in the FLAIR sequence along with manual FLAIR abnormality segmentation masks which are binary, 1-channel images considered as ground truth. The images for the data set have been obtained from The Cancer Imaging Archive (TCIA). In order to find out the brain tumor from modalities of the brain MRI images, image segmentation was performed along with some post processing steps. The representative segmentation achieved after running our proposed algorithm is depicted in Figure [3.2] and Figure [3.3] for two different patients as an example where two of the best segmented clusters generated by the algorithm are merged as a post processing step in order to compare with the ground truth.

[1] https://www.kaggle.com/mateuszbuda/lgg-mri-segmentation

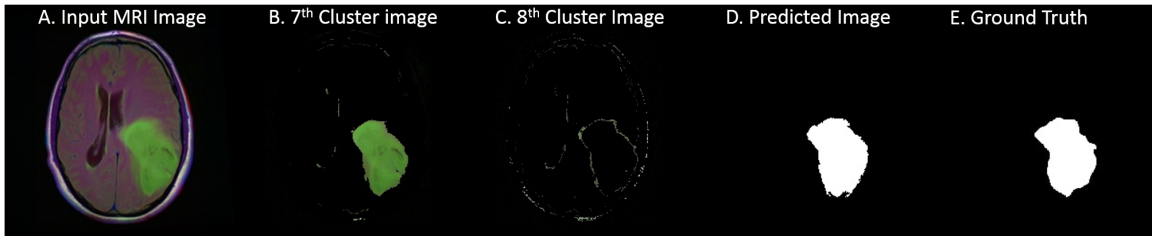

Figure 3.2: Example of best segmented brain MRI images for patient 1 : A. Input MRI image, B. 7^{th} Cluster Image, C. 8^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image

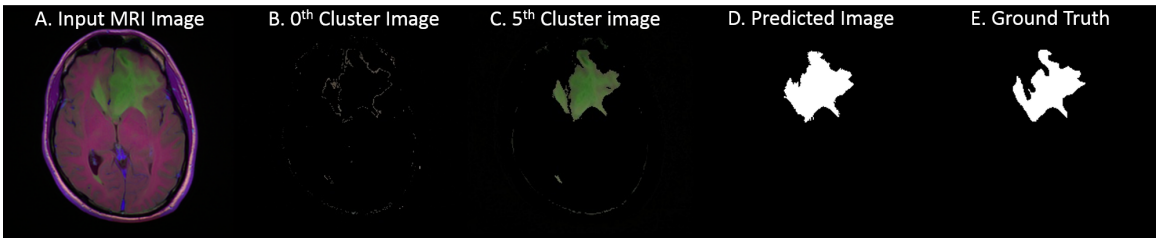

Figure 3.3: Example of best segmented brain MRI images for patient 2 : A. Input MRI image, B. 0^{th} Cluster Image, C. 5^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image.

Table [3.2] below shows the performance comparison between our proposed algorithm, OVIDMM and OVGMM . The result obtained from our algorithm are clearly much better in terms of accuracy for all evaluation metrics compared to OVIDMM and OVGMM signifying our algorithm could be of better use in healthcare to diagnose brain tumor.

Method	ARI	AMI	V-Measure	Dice	Jaccard
OVGIDMM	**90.44**	**78.66**	**80.80**	**91.12**	**82.97**
OVIDMM	84.02	67.9	72.11	86.38	75.0
OVGMM	82.3	65.83	70.90	83.63	73.02

Table 3.2: Evaluation metrics for brain tumor detection

Skin Melanoma detection

In recent years, skin cancer has emerged to be a high burden disease. Depending on the cause, there are three different consequent skin cancer conditions that can arise: melanoma, Cutaneous squamous cell carcinoma and cutaneous basal cell carcinoma, of which melanoma is the most unpredictable [91]. Melanoma accounts for 75% of all skin cancer deaths [91, 92]. Intensive skin exposure to ultraviolet radiation is the leading cause of melanoma. If diagnosed and treated in its early stages, it can be cured, but if the diagnosis is delayed, melanoma can grow deeper into the skin and spread to other parts of the body [92]. The origin of melanoma cells is unknown, it has been proposed that melanoma cells arise from either dedifferentiated melanocytes or from melanocyte progenitors [93].

Dermoscopy is a non invasive examination technique based on the use of incident light and oil immersion to enable visual examination of skin surface structures. Although detection of melanoma by dermoscopy is superior to discovery based on unaided observation, its diagnostic accuracy depends on dermatologist training, where the diagnostic accuracy of melanoma is estimated to be about 75-84%. Therefore, a lot of research is put into establishing good segmentation techniques to detect melanoma and to assist doctors in their diagnosis. Such computer aided diagnostics can improve in accuracy of melanoma detection as it can extract some information, such as color variation, asymmetry, and plot characteristics, which may not be readily apparent to human eyes. The feature extraction methodology of many computerized melanoma detection systems was primarily based on the conventional clinical algorithm of the ABCD dermoscopy rule due to its effectiveness and simplicity of implementation [94].

In order to test the performance of our algorithm to detect skin melanoma we used the data set from International Skin Imaging Collaboration[2]. The data set consists of 23,906 dermoscopic images of melanoma of different patients with ground truth available for each image. Figures [3.4] and [3.5] are example images of two different patients respectively showing the result of our proposed algorithm while performing image segmentation with feature selection. In each patient's case there were a lot of cluster images formed upto approximately the number of components however, in post processing of the image we merged the best segmented images in order to compare it with ground truth.

[2]https://www.isic-archive.com

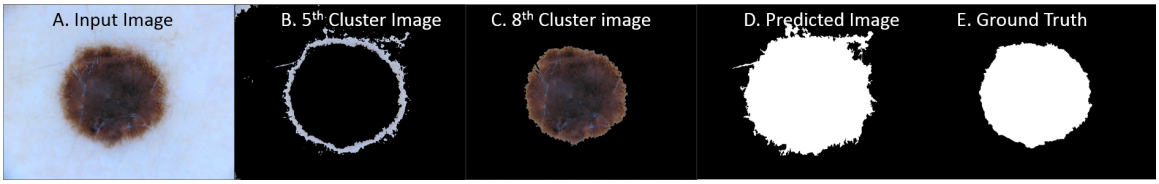

Figure 3.4: Example of best segmented dermoscopic images for patient 1 : A. Input image, B. 5^{th} Cluster Image, C. 8^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image

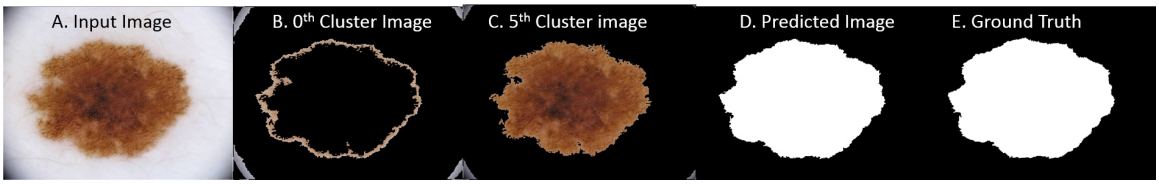

Figure 3.5: Example of best segmented dermoscopic images for patient 2 : A. Input image, B. 0^{th} Cluster Image, C. 5^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image.

Table [3.3] below shows the result obtained from our algorithm, as compared to OVIDMM and OVGMM whihc is much accurate for all evaluation metrics where we took mean of the test performed on 30 sample images.

Method	ARI	AMI	V-Measure	FM	Dice	Jaccard
OVGIDMM	**88.22**	**75.04**	**78.56**	**94.59**	**95.95**	**92.48**
OVIDMM	75.46	62.43	67.03	89.70	89.67	82.17
OVGMM	50.57	42.86	46.04	82.32	69.90	58.81

Table 3.3: Evaluation metrics for skin melanoma detection

Malaria data set

Malaria is vector borne parasitic disease prevalent in tropical parts of the world. It is transmitted by the bite of female *Anopheles* mosquito. *Plasmodium Falciparum* and *plasmodium vivax* are predominant parasite transmitted by mosquitoes worldwide with the highest occurance rate of malaria cases. About 80% of falciparum malaria cases are reported in Africa. The absolute global burden of malaria is unknown due to various factors, such as the increasing prevalence in some areas due to the wide availability of fake and substandard drugs; the expansion of drug resistance; global warming, climate change and malaria expansion in favorable regions at higher altitudes. The gold standard of laboratory diagnosis of malaria remains light microscopy of stained blood films. The blood films are stained by giemsa dye, where the trophozoites are stained in the red blood cell(RBC). Due to the staining involved and observation of the blood film under microscope, this is a complicated process that requires specialized technicians.

The infection of Malaria parasite causes micro structural changes to the erythrocytes. The RBCs microscopic features are usually specified to morphology, intensity and texture. Also, they may perform the differences that happen between healthy and unhealthy cells. Both textural and geometric merits for demonstrating stages of malaria infection have been reported in most of the studies. In general, merits may be identified according to the next characteristics: morphological features and textural and intensity features [95]. It is a popular arithmetical morphology procedure to compute the grains size distribution in binary images, by a sequence of morphological opening operations. Some authors utilize the area granulometry for prepossessing goals in malaria description, although it is certainly efficient for extracting cell size features. Local area granulometry connected with colour histogram are employed as features. The feature of area granulometry is computed locally on the stained objects binary mask, for channels of RGB.

Computer vision is a growing field for the early detection of malaria. It employs mathematical morphology as a powerful tool to develop computer aided malaria diagnosis (CAD) due to the frequent practical difficulties encountered in resource-poor health facilities in developing countries, such as an excessive workload due to lack of staff [96, 97]. Such CAD enabled identification of parasitic vs non-parasitic cells helps to reduce dependence on manual microscopic examination of blood smears, which is a thorough and time-consuming

activity while requiring considerable knowledge of the laboratory technician. Furthermore, on the identification of presence of parasite, an additional classification of the parasitic life-stage.

In this chapter, we used the malaria data set from NIH [3]. The data set includes a sum of 27,558 cell images with equivalent examples of parasitized and uninfected cells. A few examples of the images from the data set are illustrated in Figure [3.6] of parasitized cells and Figure [3.7] of uninfected cells. The data set also includes a csv file including the Patient-ID to cell mappings for the parasitized and uninfected classes. There are 151 patient entries for the parasitized class and the uninfected class includes 201 entries as the normal cells. In our experiments, the feature selection concept played a very crucial role in this data set to evaluate the performance of our algorithm. The features were extracted using the color histogram method where we considered specifically the green component of RGB model since the red and blue had no variations. The same has been described in the above section [3.3.1]. Feature extraction has the target of decreasing the subsequent computational complication and facilitating a credible and accurate recognition for unknown new data. For this experiment, the number of components M and K had been initialized with 2 and 4 and the feature salencies value were initialized at 0.5. The initial values of the hyper-parameters u, p, g, v and s for the conjugate priors are fixed to 1, q, h and t were set to 10. ϵ and Σ the learning rate parameters are fixed to 0.5 and 64. In total, we considered 17 features out of which 4 were considered as relevant and the rest as irrelevant.

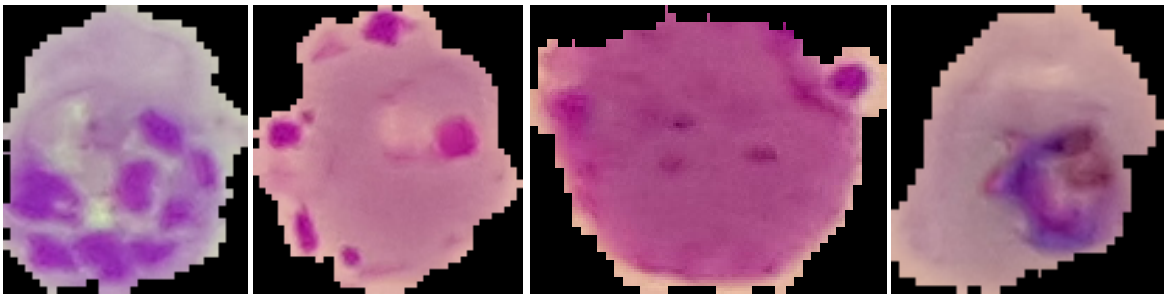

Figure 3.6: Examples of malaria cells labelled as parasitized in the data set

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

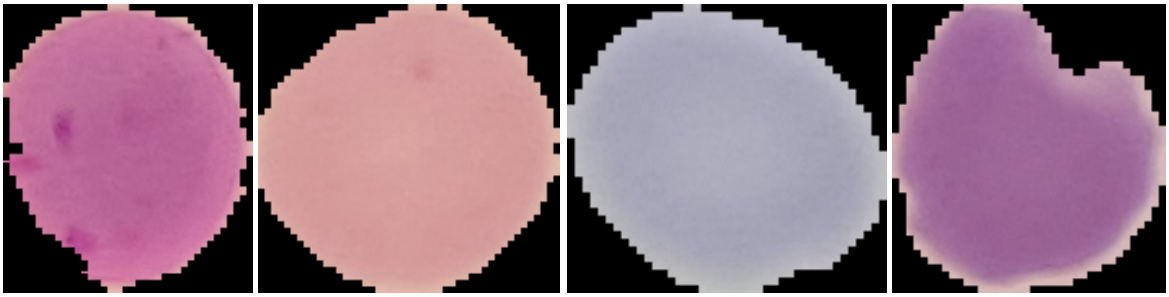

Figure 3.7: Examples of malaria cells labelled as uninfected in the data set

Table [3.4] shows precisely the way our proposed algorithm out performs the other two algorithm in CAD of malaria by giving a greater accuracy as well as taking less time for execution. It also proves the fact that it takes less time for convergence as compared to the other two models.

Method	Accuracy	Precision	Recall	F1-score	estimation time(sec)
OVGIDMM	**93.8**	**95.10**	**90.06**	**92.51**	**0.3**
OVIDMM	90.3	87.96	79.47	87.59	0.5
OVGMM	83.80	75.26	92.71	83.08	1.2

Table 3.4: Evaluation metrics for malaria data set

Chapter 4

Online Variational learning for Finite Inverted Beta-Liouville Mixture Model

In this chapter, we propose a statistical framework for online variational learning of finite inverted beta-liouville mixture model for clustering medical images data sets. We evaluated our proposed algorithm on five different biomedical image data sets including optic disc detection and localization in diabetic retinopathy, digital imaging in melanoma lesion detection and segmentation, brain tumour detection, colon cancer detection and computer aid detection (CAD) of Malaria. Furthermore, we compared the proposed algorithm with three other popular algorithms. In our results we analyse that the proposed online variational learning of finite inverted beta-liouville mixture model algorithm performs accurately on multiple modalities of medical images. We believe that the proposed algorithm has the capacity to address multi modal biomedical image data sets and can be further applied by researchers to analyse correct disease patterns.

4.1 Model Specification

4.1.1 Finite Inverted Beta-Liouville Mixture Model

Consider a D-dimensional vector $\mathbf{X}_i = (X_1, X_2, ..., X_D)$ from a set of N independent and identically distributed data samples $\chi = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$ generated from an inverted Beta-Liouville (IBL) distribution [98]. Then, the probability density function $p(X_i|\alpha_1, ..., \alpha_D, \alpha, \beta, \lambda)$

is given by:

$$p(\mathbf{X}_i|\alpha_{i1},...,\alpha_{iD},\alpha,\beta,\lambda) = \frac{\Gamma(\sum_{l=1}^{D}\alpha_l)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{l=1}^{D}\frac{X_{il}^{\alpha_l-1}}{\Gamma(\alpha_l)}$$
$$\times\lambda^{\beta}\left(\sum_{l=1}^{D}X_{il}\right)^{\alpha-\sum_{l=1}^{D}\alpha_l}\left(\lambda+\sum_{l=1}^{D}X_{il}\right)^{-(\alpha+\beta)} \tag{4.1}$$

with the conditions $X_{il} > 0$ for $l = 1,....,D, \alpha > 0, \beta > 0$ and $\lambda > 0$. Examples of IBL mixture model is shown in Figure 4.1. The mean, variance and covariance of IBL distribution are given by:

$$E(X_{il}) = \frac{\lambda\alpha}{\beta-1}\frac{\alpha_l}{\sum_{l=1}^{D}\alpha_l} \tag{4.2}$$

$$Var(X_{il}) = \frac{\lambda^2\alpha(\alpha+1)}{(\beta-1)(\beta-2)}\frac{\alpha(\alpha+1)}{\alpha_l(\sum_{l=1}^{D}\alpha_l+1)}\frac{\lambda^2\alpha^2}{(\beta-1)^2}\frac{\alpha_l^4}{(\sum_{l=1}^{D}\alpha_l)^4} \tag{4.3}$$

$$Cov(X_{im},X_{in}) = \frac{\alpha_m\alpha_n}{\sum_{l=1}^{D}\alpha_l}\left[\frac{\lambda^2\alpha(\alpha+1)}{(\beta-1)(\beta-2)(\sum_{l=1}^{D}\alpha_l+1)} - \frac{\lambda^2\alpha^2}{(\beta-1)^2(\sum_{l=1}^{D}\alpha_l)}\right] \tag{4.4}$$

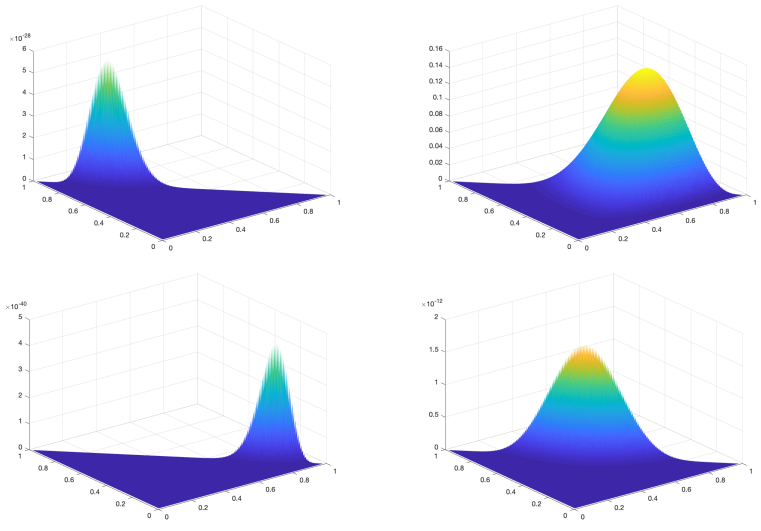

Figure 4.1: Four examples of inverted beta liouville distributions

Let's assume that given a set of data that contains N vectors where each sample $\vec{X}_i$ = $(X_{i1}, X_{i2}, ..., X_{iD})$ is generated from a mixture of IBL distributions then :

$$p(\vec{X}_i|\vec{\pi}, \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{M} \pi_j p(\vec{X}_i|\theta_j) \tag{4.5}$$

where *M* is the number of components in the mixture model and $\Theta = (\theta_1, \theta_2, ..., \theta_{\boldsymbol{M}})$, $p(\vec{X}_i|\theta_j)$ denotes the conditional probability of the data sample with respect to each component, $\theta_j = (\alpha_{j1}, ..., \alpha_{jD}, \alpha_j, \beta_j, \lambda_j)$ represents the parameter with respect to the component j . $\vec{\pi} = (\pi_1, ..., \pi_{\boldsymbol{M}})$ is the set of mixing parameters and follows the conditions $\sum_{j=1}^{M} \pi_j = 1$ and $0 \leq \pi_j \leq 1$. Examples of our mixture model with different components is represented in Figure 4.2.

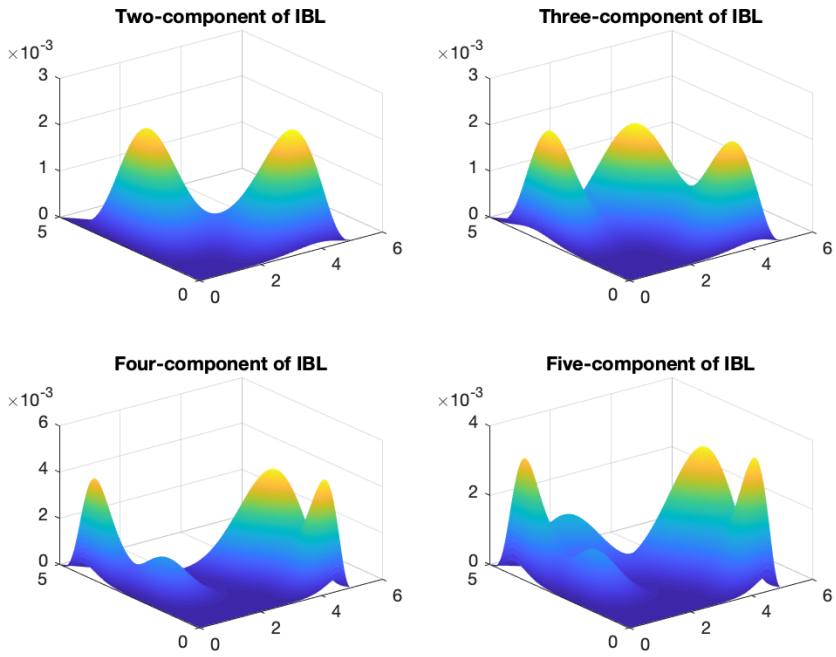

Figure 4.2: Examples of finite IBL Mixture model with different components

We define latent variables $\boldsymbol{\mathcal{Z}} = (\vec{Z}_1, ..., \vec{Z}_N)$ as an indicator matrix which indicates to which component each data sample is assigned to [99]. Here each $\vec{\mathcal{Z}}_i = (\mathbf{Z}_{i1}, ..., \mathbf{Z}_{i\boldsymbol{M}})$. $\vec{Z}_i$

is a binary vector that satisfies the conditions $Z_{ij}\epsilon\{0,1\}$ and $\sum_{j=1}^{M} Z_{ij} = 1$ and is defined by:

$$Z_{ij} = \begin{cases} 1, & if \mathbf{X}_i \epsilon j \\ 0, & otherwise. \end{cases}$$

The conditional distribution of $\mathcal{Z}$ can be defined as:

$$p(\mathcal{Z}|\vec{\pi}) = \prod_{i=1}^{N}\prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{4.6}$$

Therefore, according to equation (4.6), the likelihood function of data set χ with latent variables $\mathcal{Z}$ and related parameters Θ is given by as:

$$p(\chi|\mathcal{Z},\Theta) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(\mathbf{X}_i|\theta_j)^{Z_{ij}} \tag{4.7}$$

We now place priors over the parameters $\Theta = (\alpha, \overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\lambda})$. Since all the parameters are positive thus we make the choice of modelling them using Gamma prior. Hence the priors are defined by:

$$p(\alpha_{jl}) = \mathcal{G}(\overrightarrow{\alpha} \mid \overrightarrow{u}, \overrightarrow{\nu}) = \prod_{j=1}^{M}\prod_{d=1}^{D} \frac{\nu_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{4.8}$$

$$p(\alpha_j) = \mathcal{G}(\overrightarrow{\alpha} \mid \overrightarrow{p}, \overrightarrow{q}) = \prod_{j=1}^{M} \frac{q_j^{p_j}}{\Gamma(p_j)} \alpha_j^{p_j-1} e^{-q_j\alpha_j} \tag{4.9}$$

$$p(\beta_j) = \mathcal{G}(\overrightarrow{\beta} \mid \overrightarrow{g}, \overrightarrow{h}) = \prod_{j=1}^{M} \frac{h_j^{g_j}}{\Gamma(g_j)} \beta_j^{g_j-1} e^{-h_j\beta_j} \tag{4.10}$$

$$p(\lambda_j) = \mathcal{G}(\overrightarrow{\lambda} \mid \overrightarrow{s}, \overrightarrow{t}) = \prod_{j=1}^{M} \frac{t_j^{s_j}}{\Gamma(s_j)} \lambda_j^{s_j-1} e^{-t_j\lambda_j} \tag{4.11}$$

where all the hyper-parameters $\overrightarrow{u} = \{u_{jl}\}$, $\overrightarrow{v} = \{v_{jl}\}$, $\overrightarrow{p} = \{p_j\}$, $\overrightarrow{q} = \{q_j\}$, $\overrightarrow{g} = \{g_j\}$, $\overrightarrow{h} = \{h_j\}$, $\overrightarrow{s} = \{s_j\}$ and $\overrightarrow{t} = \{t_j\}$ of the above conjugate priors are positive. Therefore, the joint distribution of all random variables and latent variables given mixing coefficient π is defined by

$$
\begin{aligned}
p(\boldsymbol{\chi}, \mathbf{Z}, \boldsymbol{\Theta}|\overrightarrow{\boldsymbol{\pi}}) =& p(\boldsymbol{\chi}|\mathbf{Z}, \boldsymbol{\Theta}) p(Z|\overrightarrow{\boldsymbol{\pi}}) p(\alpha) p(\overrightarrow{\boldsymbol{\alpha}}) p(\overrightarrow{\boldsymbol{\beta}}) p(\overrightarrow{\boldsymbol{\lambda}}) \\
=& \prod_{i=i}^{N} \prod_{j=i}^{M} \left[\frac{\Gamma(\sum_{l=1}^{D} \alpha_{jl}) \Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \prod_{l=1}^{D} \frac{X_{il}^{\alpha_{il}-1}}{\Gamma(\alpha_{jl})} \right. \\
& \times \lambda_j^{\beta_j} \left(\sum_{l=1}^{D} X_{il} \right)^{\alpha_j - \sum_{i=1}^{D} \alpha_{jl}} \left(\lambda_j + \sum_{l=1}^{D} X_{il} \right)^{-(\alpha_j + \beta_j)} \Bigg]^{Z_{ij}} \\
& \times \prod_{i=1}^{N} \left[\prod_{j=1}^{s} \pi_j^{Z_{ij}} \right] \times \prod_{j=1}^{M} \prod_{l=1}^{D} \left[\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-v_{jl}\alpha_{jl}} \right. \\
& \left. \times \frac{q_j^{p_j}}{\Gamma(p_j)} \alpha_j^{p_j - 1} e^{-q_j \alpha_j} \times \frac{h_j^{g_j}}{\Gamma(g_j)} \beta_j^{g_j - 1} e^{-h_j \beta_j} \times \frac{t_j^{s_j}}{\Gamma(s_j)} \lambda_j^{s_j - 1} e^{-t_j \lambda_j} \right]
\end{aligned}
\tag{4.12}
$$

4.2 Online variational learning for finite Inverted Beta-Liouville Mixture Model

In the past decades variational procedures have been extensively utilized and commonly used to find approximations which are tractable for the posterior distributions of a variety of statistical models [74]. In this chapter we take into consideration the online variational framework of finite IBL mixture model for parameter estimation and model selection. The concept of variational inference has been explained in detail in the previous chapters. The graphical representation of the finite IBL mixture model is in Figure 4.3.

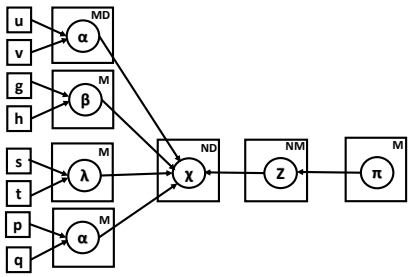

Figure 4.3: Graphical model representation for finite IBL mixture model. Symbols in the circle denote the random variables; otherwise, they denote the model parameters.

The parametric forms of the variational posterior distributions could be defined as following:

$$\boldsymbol{\mathcal{Q}}(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{4.13}$$

$$\boldsymbol{\mathcal{Q}}(\alpha) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^{*}\nu_{jl}^{*}), \qquad \boldsymbol{\mathcal{Q}}(\alpha) = \prod_{j=1}^{M} \mathcal{G}(\alpha_{j}|p_{j}^{*}, q_{j}^{*}) \tag{4.14}$$

$$\boldsymbol{\mathcal{Q}}(\beta) = \prod_{j=1}^{M} \mathcal{G}(\beta_{j}|g_{j}^{*}, h_{j}^{*}), \qquad \boldsymbol{\mathcal{Q}}(\lambda) = \prod_{j=1}^{M} \mathcal{G}(\lambda_{j}|s_{j}^{*}, t_{j}^{*}) \tag{4.15}$$

where:

$$r_{ij} = \frac{r_{ij}^{*}}{\sum_{k=1}^{M} r_{ik}^{*}} \tag{4.16}$$

$$\begin{aligned} r_{ij}^{*} = exp\{ln\pi_{j} + \boldsymbol{R}_{j} + \boldsymbol{S}_{j} + (\bar{\alpha}_{j} - \sum_{l=1}^{D}\bar{\alpha}_{jl})ln(\sum_{l=1}^{D} X_{il}) + \bar{\beta}_{j}\langle ln\lambda_{j}\rangle \\ + \sum_{l=1}^{D}[(\bar{\alpha}_{jd} - 1)lnX_{id}] - (\bar{\alpha} + \bar{\beta})T_{ij}\} \end{aligned} \tag{4.17}$$

where R_j , S_j and T_{ij} are given by equations (4.18, 4.19, 4.20) as below :

$$\begin{aligned} \boldsymbol{R}_{j} = & ln\frac{\Gamma(\sum_{l=1}^{D}\bar{\alpha}_{jl})}{\prod_{l=1}^{D}\Gamma(\bar{\alpha_{jl}})} + \sum_{l=1}^{D}\bar{\alpha}_{jl}\left[\psi\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right) - \psi\left(\bar{\alpha}_{jl}\right)\right]\left[\langle ln\alpha_{jl}\rangle - ln\bar{\alpha}_{jl}\right] \\ & + \frac{1}{2}\sum_{l=1}^{D}\bar{\alpha}_{jl}^{2}\left[\psi'\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right) - \psi'\left(\bar{\alpha}_{jl}\right)\right]\left\langle\left(ln\alpha_{jl} - ln\bar{\alpha}_{jl}\right)^{2}\right\rangle \\ & + \frac{1}{2}\sum_{a=1}^{D}\sum_{b=1}^{D}\bar{\alpha}_{ja}\bar{\alpha}_{jb}\left[\psi'\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right)\left(\langle ln\alpha_{ja}\rangle - ln\bar{\alpha}_{ja}\right)\right. \\ & \left.\times\left(\langle ln\alpha_{jb}\rangle - ln\bar{\alpha}_{jb}\right)\right] \end{aligned} \tag{4.18}$$

$$
\begin{aligned}
S_j = \ln \frac{\Gamma\big(\overline{\alpha}+\overline{\beta}\big)}{\Gamma\big(\overline{\beta}\big)\Gamma\big(\overline{\alpha}\big)} \\
&+ \overline{\alpha}\Big[\psi\big(\overline{\alpha}+\overline{\beta}\big) - \psi\big(\overline{\alpha}\big)\Big]\Big[\big\langle \ln\alpha\big\rangle - \ln\overline{\alpha}\Big] \\
&+ \overline{\beta}\Big[\psi\big(\overline{\beta}+\overline{\alpha}\big) - \psi\big(\overline{\beta}\big)\Big]\Big[\big\langle \ln\beta\big\rangle - \ln\overline{\beta}\Big] \\
&+ 0.5\overline{\alpha}^2\Big[\psi'\big(\overline{\alpha}+\overline{\beta}\big) - \psi'\big(\overline{\alpha}\big)\Big]\Big[\big\langle \ln\alpha - \ln\overline{\alpha}\big\rangle\Big]^2 \\
&+ 0.5\overline{\beta}^2\Big[\psi'\big(\overline{\beta}+\overline{\alpha}\big) - \psi'\big(\overline{\beta}\big)\Big]\Big[\big\langle \ln\beta - \ln\overline{\beta}\big\rangle\Big]^2 \\
&+ \overline{\alpha}\overline{\beta}\psi\big(\overline{\alpha}+\overline{\beta}\big)\Big[\big\langle \ln\beta\big\rangle - \ln\overline{\beta}\Big]\Big[\big\langle \ln\alpha\big\rangle - \ln\overline{\alpha}\Big]
\end{aligned} \tag{4.19}
$$

$$
T_{ij} = ln\left[\bar{\lambda}_j + \sum_{l=1}^{D} X_{il}\right] + \frac{\bar{\lambda}_j}{\bar{\lambda}_j + \sum_{i=1}^{D} X_{il}}\left[\langle ln\lambda_j - ln\bar{\lambda}_j\right] \tag{4.20}
$$

$$
u^*_{jl} = u_{jl} + \sum_{i=1}^{N}\langle Z_{ij}\rangle\bar{\alpha}_{jl}\left[\psi(\sum_{l=1}^{D}\bar{\alpha}_{jl}) - \psi(\bar{\alpha}_{jl}) + \psi\prime(\sum_{l=1}^{D}\bar{\alpha}_{jl})\sum_{d\neq l}^{D}(\langle ln\alpha_{jd}\rangle - ln\bar{\alpha}_{jd})\bar{\alpha}_{jd}\right] \tag{4.21}
$$

$$
v^*_{jl} = v_{jl} - \sum_{i=1}^{N}\langle Z_{ij}\rangle\left[lnX_{il} - ln(\sum_{l=1}^{D} X_{il})\right] \tag{4.22}
$$

$$
p^*_j = p_j + \sum_{i=1}^{N}\langle Z_{ij}\rangle\left[\psi(\bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}) - \psi(\bar{\boldsymbol{\alpha}}_j) + \bar{\boldsymbol{\beta}}_j\psi'(\bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}_j)\Big(\langle ln\beta_j\rangle - \bar{\beta}_j\Big)\right]\bar{\boldsymbol{\alpha}}_j \tag{4.23}
$$

$$
q^*_j = q_j - \sum_{i=1}^{N}\langle Z_{ij}\rangle ln(\sum_{l=1}^{D} X_{il}) + \sum_{i=1}^{N}\langle Z_{ij}\rangle T_{ij} \tag{4.24}
$$

$$
g^*_j = g_j + \sum_{i=1}^{N}\langle Z_{ij}\rangle\left[\psi(\bar{\alpha}_j + \bar{\beta}) - \psi(\bar{\beta}_j) + \bar{\alpha}_j\psi'(\bar{\alpha}_j + \bar{\beta}_j)\Big(\langle ln\alpha_j\rangle - \bar{\alpha}_j\Big)\right]\bar{\beta}_j \tag{4.25}
$$

$$
h^*_j = h_j + \sum_{i=1}^{N}\langle Z_{ij}\rangle\Big[T_{ij} - \langle ln\lambda_j\rangle\Big] \tag{4.26}
$$

$$
s^*_j = s_j + \sum_{i=1}^{N}\langle Z_{ij}\rangle\bar{\beta}_j \tag{4.27}
$$

$$t_j^* = t_j + \sum_{i=1}^{N} \langle Z_{ij} \rangle \frac{\bar{\alpha}_j + \bar{\beta}_j}{\bar{\lambda}_j + \sum_{l=1}^{D} X_{il}} \tag{4.28}$$

The first and second derivative of the Gamma function is given by the digamma and trigamma functions, $\psi(\cdot)$ and $\psi'(\cdot)$ respectively. The values of the expectations mentioned in the above equations are given by:

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

$$\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}}{v_{jl}}, \quad \bar{\alpha}_j = \langle \alpha_j \rangle = \frac{p_j}{q_j}, \quad \bar{\beta}_j = \langle \beta_j \rangle = \frac{g_j}{h_j}, \quad \bar{\lambda}_j = \langle \lambda_j \rangle = \frac{s_j}{t_j} \tag{4.29}$$

$$\langle ln\alpha_{jl} \rangle = \psi(u_{jl}^*) - lnv_{jl}^*, \quad \langle ln\alpha_j \rangle = \psi(p_j^*) - lnq_j^*, \tag{4.30}$$

$$\langle ln\beta_j \rangle = \psi(g_j^*) - lnh_j^*, \quad \langle ln\lambda_j \rangle = \psi(s_j^*) - lnt_j^* \tag{4.31}$$

$$\langle (ln\alpha_{jl} - ln\bar{\alpha}_{jl})^2 \rangle = \left[\psi(u_{jl}^*) - lnu_{jl}^*\right]^2 + \psi'(u_{jl}^*) \tag{4.32}$$

$$\langle (ln\alpha_j - ln\bar{\alpha}_j)^2 \rangle = \left[\psi(p_j^*) - lnp_j^*\right]^2 + \psi'(p_j^*) \tag{4.33}$$

$$\langle (ln\beta_j - ln\bar{\beta}_j)^2 \rangle = \left[\psi(g_j^*) - lng_j^*\right]^2 + \psi'(g_j^*) \tag{4.34}$$

We propose an online variational framework for finite IBL mixture model by adopting the framework proposed in [77]. In our case, let t denotes the actual amount of observed data. Then, the current lower bound for the observed data is give by

$$\boldsymbol{\mathcal{L}}^{(t)}(\boldsymbol{\mathcal{Q}}) = \frac{N}{t} \sum_{i=1}^{t} \int \boldsymbol{\mathcal{Q}}(\Lambda) d\Lambda \sum_{\vec{Z}_i} \boldsymbol{\mathcal{Q}}(\vec{Z}_i) ln \left[\frac{p(\vec{X}_i, \vec{Z}_i | \Lambda)}{\mathcal{Q}(\vec{Z}_i)} \right] + \int \boldsymbol{\mathcal{Q}}(\Lambda) ln \left[\frac{p(\Lambda)}{\boldsymbol{\mathcal{Q}}(\Lambda)} \right] d\Lambda \tag{4.35}$$

where $\Lambda = \{\vec{\pi}, \vec{\theta}\}$. The key idea of the online variational learning algorithm is to successively maximize the current variational lower bound in equation (4.35). Assume that we have already observed a data set $\{X_1, ...s, X_{(t-1)}\}$. For a new observation X_t, we can maximize the current lower bound $\boldsymbol{\mathcal{L}}^{(t)}(\boldsymbol{\mathcal{Q}})$ with respect to $\mathcal{Q}(\vec{\mathcal{Z}}_t)$ while other variational

factors are fixed to $\mathcal{Q}^{(t-1)}(\vec{\lambda})$, $\mathcal{Q}^{(t-1)}(\vec{\boldsymbol{\alpha}})$, $\mathcal{Q}^{(t-1)}(\vec{\alpha})$, $\mathcal{Q}^{(t-1)}(\vec{\beta})$ and $\mathcal{Q}^{(t-1)}(\vec{\pi})$. Thus, the variational solution to $\boldsymbol{\mathcal{Q}}(\boldsymbol{Z})^{(t)}$ is given by

$$\boldsymbol{\mathcal{Q}}(\boldsymbol{Z})^{(t)} = \prod_{j=1}^{M} r_{tj}^{Z_{tj}} \tag{4.36}$$

where

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

where we substitute equation (4.17) and it becomes as below for online case

$$\tilde{r}_{tj} = exp\{ln\pi_j^{(t-1)} + R_j^{(t-1)} + S_j^{(t-1)} + (\bar{\alpha}_j^{(t-1)} - \sum_{l=1}^{D} \bar{\alpha}_{jl}^{(t-1)}) ln(\sum_{l=1}^{D} X_{il}) + \bar{\beta}_j^{(t-1)} \langle ln\lambda_j^{(t-1)} \rangle \tag{4.37}$$

$$+ \sum_{l=1}^{D} \Big[(\bar{\alpha}_{jl}^{(t-1)} - 1) lnX_{il} \Big] - (\bar{\alpha}^{(t-1)} + \bar{\beta}^{(t-1)}) T_{ij}^{(t-1)} \}$$

Next, the current lower bound $\boldsymbol{\mathcal{L}}^{(t)}(\boldsymbol{\mathcal{Q}})$ is maximized with respect to $\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\lambda})$, while $\boldsymbol{\mathcal{Q}}(\boldsymbol{Z})^{(t)}$ is fixed and other variational factors remain at their $(t-1)^{th}$ values. Therefore, we can obtain the variational solution to $\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\lambda})$ as

$$\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\lambda}) = \prod_{j=1}^{M} \mathcal{G}(\lambda_j^{(t)} | s_j^{(t)}, t_j^{(t)}) \tag{4.38}$$

A significant characteristic of the adopted variational method [52], which cites that variational inference could be handled as a natural gradient method [56] has been desrcibed in detail in earlier chapters. Here , $\Delta s_j^{(t)}$ and $\Delta t_j^{(t)}$ are the natural gradients of the corresponding hyper parameters which are given by :

$$\Delta s_j^{*(t)} = s_j^{*(t)} - s_j^{*(t-1)} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \bar{\beta}_j \tag{4.39}$$

$$\Delta t_j^{*(t)} = t_j^{*(t)} - t_j^{*(t-1)} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \frac{\bar{\alpha}_j + \bar{\beta}_j}{\bar{\lambda}_j + \sum_{l=1}^{D} X_{il}} \tag{4.40}$$

where the hyper parameters are defined by

$$\begin{aligned} s_j^{(t)} &= s_j^{(t-1)} + \rho_t \Delta s_j^{(t)}, \\ t_j^{(t)} &= t_j^{(t-1)} + \rho_t \Delta t_j^{(t)} \end{aligned} \tag{4.41}$$

where ρ_t is the learning rate and is defined as

$$\rho_t = (\eta_o + t)^{-a} \tag{4.42}$$

In this chapter, we adopt a learning rate function introduced in [57] which is in equation (4.42) subject to the constraints $a\epsilon(0.5, 1)$ and $\eta_0 \geq 0$.

Subsequently, the current lower bound $\boldsymbol{\mathcal{L}}^{(t)}(\boldsymbol{\mathcal{Q}})$ is maximized with respect to $\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\alpha}_l)$, and the corresponding variational solution is given by

$$\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\alpha}_l) = \prod_{j=1}^{M}\prod_{l=1}^{D} \boldsymbol{\mathcal{G}}(\alpha_{jl}^{(t)}|u_{jl}^{*(t)}, v_{jl}^{*(t)}) \tag{4.43}$$

where

$$\begin{aligned} u_{jl}^{*(t)} &= u_{jl}^{*(t-1)} + \rho\Delta u_{jl}^{*(t)}, \\ v_{jl}^{*(t)} &= v_{jl}^{*(t-1)} + \rho\Delta v_{jl}^{*(t)} \end{aligned} \tag{4.44}$$

The corresponding natural gradients are defined by

$$\Delta u_{jl}^{*(t)} = u_{jl}^{*(t)} - u_{jl}^{*(t-1)} = \sum_{i=1}^{N}\langle Z_{ij}\rangle\bar{\alpha}_{jl}\left[\psi(\sum_{l=1}^{D}\bar{\alpha}_{jl}) - \psi(\bar{\alpha}_{jl}) + \psi\prime(\sum_{l=1}^{D}\bar{\alpha}_{jl})\sum_{d\neq l}^{D}(\langle ln\alpha_{jd}\rangle - ln\bar{\alpha}_{jd})\bar{\alpha}_{jd}\right] \tag{4.45}$$

$$\Delta v_{jl}^{*(t)} = v_{jl}^{*(t)} - v_{jl}^{*(t-1)} = -\sum_{i=1}^{N}\langle Z_{ij}\rangle\left[lnX_{il} - ln(\sum_{l=1}^{D}X_{il})\right] \tag{4.46}$$

The hyper parameters solutions of $Q^{(t)}(\vec{\alpha})$ and $Q^{(t)}(\vec{\beta})$ can be calculated similarly. In order to do that the current lower bound $\boldsymbol{\mathcal{L}}^{(t)}(\boldsymbol{\mathcal{Q}})$ is maximized with respect to $\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\alpha})$ and $\boldsymbol{\mathcal{Q}}^{(t)}(\vec{\beta})$ and the corresponding variational solutions are given by

$$\boldsymbol{\mathcal{Q}}^{(\boldsymbol{t})}(\alpha) = \prod_{j=1}^{M}\mathcal{G}(\alpha_j^{(t)}|p_j^{*(t)}, q_j^{*(t)}) \tag{4.47}$$

$$\boldsymbol{\mathcal{Q}}^{(\boldsymbol{t})}(\beta) = \prod_{j=1}^{M}\mathcal{G}(\beta_j^{(t)}|g_j^{*(t)}, h_j^{*(t)}) \tag{4.48}$$

where

$$p_j^{*(t)} = p_j^{*(t-1)} + \rho\Delta p_j^{*(t)} \tag{4.49}$$

$$q_j^{*(t)} = q_j^{*(t-1)} + \rho \Delta q_j^{*(t)} \tag{4.50}$$

$$g_j^{*(t)} = g_j^{*(t-1)} + \rho \Delta g_j^{*(t)} \tag{4.51}$$

$$h_j^{*(t)} = h_j^{*(t-1)} + \rho \Delta h_j^{*(t)} \tag{4.52}$$

The corresponding natural gradients for the equations (4.47) and (3.19) are give as below :

$$\Delta p_j^{*(t)} = p_j^{*(t)} - p_j^{*(t-1)} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\psi(\bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}) - \psi(\bar{\boldsymbol{\alpha}}_j) + \bar{\boldsymbol{\beta}}_j \psi'(\bar{\boldsymbol{\alpha}}_j) + \bar{\boldsymbol{\beta}}_j) \left(\langle ln \beta_j \rangle - \bar{\beta}_j \right) \right] \bar{\boldsymbol{\alpha}}_j \tag{4.53}$$

$$\Delta q_j^{*(t)} = q_j^{*(t)} - q_j^{*(t-1)} = - \sum_{i=1}^{N} \langle Z_{ij} \rangle ln \left(\sum_{l=1}^{D} X_{il} \right) + \sum_{i=1}^{N} \langle Z_{ij} \rangle T_{ij} \tag{4.54}$$

$$\Delta g_j^{*(t)} = g_j^{*(t)} - g_j^{*(t-1)} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\psi(\bar{\alpha}_j + \bar{\beta}) - \psi(\bar{\beta}_j) + \bar{\alpha}_j \psi'(\bar{\alpha}_j + \bar{\beta}_j) \left(\langle ln \alpha_j \rangle - \bar{\alpha}_j \right) \right] \bar{\beta}_j \tag{4.55}$$

$$\Delta h_j^{*(t)} = h_j^{*(t)} - h_j^{*(t-1)} = \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[T_{ij} - \langle ln \lambda_j \rangle \right] \tag{4.56}$$

The online variational inference for finite IBL mixture model is summarized in Algorithm 1.

Algorithm 1 Online Variational learning of the finite IBL mixture model

1. Choose the initial number of components M.
2. Initialize the values for hyper-parameters u_{jl}, ν_{jl}, p_j, q_j, g_j, h_j, s_j, t_j.
3. Using K-means algorithm, initialize the values of r_{ij}.
4. **for** $t = 1 \rightarrow N$ **do**

    i The variational E-step:

ii Update the variational solution for $\mathcal{Q}(\overrightarrow{\mathcal{Z}}_t)$ through equation (4.36)

iii The variational M-step:

iv Compute learning rate $\rho_t = (\eta_o + t)^{-a}$ as in equation (4.42)

v Calculate the natural gradients $\Delta s_j^{*(t)}$, $\Delta t_j^{*(t)}$, $\Delta u_{jl}^{*(t)}$, $\Delta \nu_{jl}^{*(t)}$, $\Delta p_j^{*(t)}$, $\Delta q_j^{*(t)}$, $\Delta g_j^{*(t)}$ and $\Delta h_j^{*(t)}$ using equations (4.39), (4.40) (4.45), (4.46), (4.53), (4.54), (4.55) and (4.56) respectively

vi Update the variational solution for $\mathcal{Q}^{(t)}(\overrightarrow{\lambda})$, $\mathcal{Q}^{t}(\overrightarrow{\alpha_l})$, $\mathcal{Q}^{(t)}(\overrightarrow{\alpha})$, $\mathcal{Q}^{(t)}(\overrightarrow{\beta})$ through equations (4.38), (4.43), (4.47), (4.48).

vii Repeat the variational E-step and M-step until new data is observed.

5. **end for**

4.3 Experimental Results

In this section, we investigate the efficiency of our proposed online variational IBL mixture model by validating it on five challenging biomedical applications by performing image segmentation and feature extraction for analysis of diseases. The biomedical data sets were chosen as the focus of this work to access the different analytical and statistical mixture model algorithms that can contribute to medicine and help identify precise diagnosis. Amongst the five data sets chosen, we performed medical image segmentation on four different data sets to identify the relevant diseases. In the last data set for malaria disease we worked on feature extraction methodology to classify the images of the data set into uninfected and parasitized category. In this work, we have used different image modalities as the input form of image data sets, i.e., magnetic resonance imaging (MRI) with FLAIR, computer aid detection (CAD), regular camera image format and microscopy image. These image modalities test the applicability of the algorithm on varied output formats. Further to this, we applied the algorithm to detect brain tumour, skin melanoma lesion, colon cancer, diabtetic retinopathy and malaria.

In order to validate the accuracy and to illustrate the merit of OVIBLMM algorithm, we compared it to the implementation of three other algorithms, namely online variational learning of finite generalized inverted dirichlet mixture model (OVGIDMM), OVIDMM and OVGMM. The below sections would follow the description on the results obtained by

calculating different evaluation metrics, namely, adjusted rand index (ARI) score, adjusted mutual information (AMI) score, V-measure score, dice similarity coefficient and Jaccard similarity index for evaluation of the accuracy on each medical data set. In the case of CAD data set of malaria, the evaluation metrics were confusion matrix, precision, recall and F1-Score for classification of the patients into infected and healthy.

In our experiments we did initialisation of the components and hyper parameters as per two categories. The image segmentation component initialisation and hyper parameters definition have been assigned in first category. In the first category, for brain tumour detection [4.3.1], optic disc detection [4.3.2], colon cancer detection [4.3.4] and skin melanoma image sections [4.3.3], the number of components M have been initialized 16. The initial values of the hyper-parameters u, p, g and s for the conjugate priors are fixed to 1 and v, q, s, t to 0.01. The learning rate parameters a and η_o are fixed to 0.5 and 64, respectively. In the second category for image clustering, different initialization for the testing of CAD of malaria data set was defined, which has been described in section [4.3.5].

4.3.1 Brain Tumor Detection

Here, we use the brain tumor magnetic resonance images(MRI) with FLAIR data set from kaggle [1]. The images were taken from The Cancer Imaging Archive (TCIA). The data set is of 110 patients suffering from low grade glioma. Furthermore, the data set has multilevel information. It codes for the patho-physiology in the images and for the cellular pathology the genomic structures are available.

Our proposed algorithm for the detection of brain tumors based on magnetic resonance imaging had higher accuracy and lower error rates. Statistical analysis of the experimental results showed that the developed algorithm can segment the brain MRI images with good precision. The representative segmentation achieved after running our proposed algorithm is depicted in Figure 4.4 for a patient as an example where two of the best segmented clusters generated by the algorithm are merged as a post processing step in order to compare with the ground truth.

[1] https://www.kaggle.com/mateuszbuda/lgg-mri-segmentation

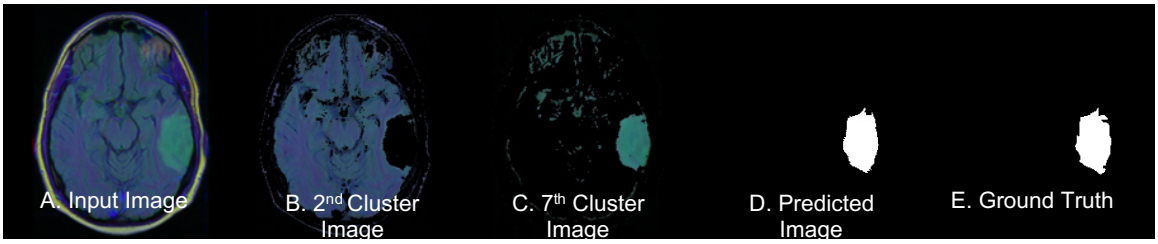

Figure 4.4: Example of best segmented brain MRI images for LGG. In the full panel from left to right are : A. Input MRI image, B. 2^{nd} Cluster Image, C. 7^{th} Cluster Image, D. Predicted Image after post processing, E. Ground Truth Image. The dice coefficient for this example was 92% The images show FLAIR modality and predict the FLAIR abnormality.

Additionally, we evaluated the performance of the proposed algorithm in comparison with other algorithms. Table 4.1 shows that the performance obtained from **OVIBLMM** algorithm as compared to OVGIDMM, OVIDMM and OVGMM is higher in terms of accuracy for all evaluation metrics.

Method	ARI	AMI	V-Measure	Dice	Jaccard
OVIBLMM	**91.82**	**80.51**	**82.66**	**92.47**	**86.26**
OVGIDMM	88.82	75.55	78.57	89.72	80.52
OVIDMM	84	67.9	72.12	86.38	75.0
OVGMM	82.26	65.83	70.91	83.63	73.02

Table 4.1: Evaluation metrics for brain tumor detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were ARI, AMI, V-measure, Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy.

4.3.2 Diabetic retinopathy (DR) Optic Disc Localization and Detection

The global burden of diabetes mellitus is very high. A total of 282 million people were estimated to be suffering from diabetes in 2013 [100]. Diabetic retinopathy (DR), in return is a consequent complication for diabetes mellitus. It remains the leading cause of visual loss in the diabetic population [101]. DR is a microvascular complication where the damaged blood vessels in the retina result in microvasular changes in the retina, which trigger vision impairment [102, 103]. The loss of vision due to DR is called diabetic macular edema (DME) which is characterised by high pressure in the blood vessels of the eye and leak of fluid trigger by the breakdown of blood retina barrier. DR falls in 2 broad categories: non proliferative diabetic retinopathy (NPDR) and PDR where the former is early stage and later is the advanced stage. NPDR can be classified with visible features such as retinal heammorages, intra retinal microvasularisation abnormalities etc. These visible features are the basis of diagnosis and detection, where emerging imaging technologies are applied and the algorithms are used to classify these features of the retinal image [104].

Optic Disc (OD) is a bright yellowish disk in human retina from where the blood vessels and optic nerves emerge. Automated localization and detection of the OD is an essential step in the analysis of digital diabetic retinopathy images. Accurate localization and detection of optic disc boundary is important for detection of PDR where fragile vessels develop in the retina [105]. Reference to diabetic patients is therefore made through regular consultation and annual or biannual monitoring to refine their retina. Eliminating the lack of justifiable views depends on the number of medical specialists and the health infrastructure needed to treat the eyes. Currently, the assessment of DR is carried out on retina fundus images by retinal experts or trained graders leading to large proportions of patients left undiagnosed due to low adherence and limited access to retina evaluation centres. Thus, in-person examination is impractical due to the size of the population suffering from DR [106]. Therefore, computer-assisted diagnostic could address the above mentioned shortcomings and help in DR management in an automated way thus reducing the labour force and allowing the diagnosis to be more accessible.

In this chapter, the data set on which the algorithms were applied is Indian Diabetic

Retinopathy Image data set (IDRiD) from [2] which provides expert markups of typical diabetic retinopathy lesions and normal retinal structures. The data set consists of 81 color fundus images with signs of DR. Precise pixel level annotation of abnormalities associated with DR like microaneurysms (MA), soft exudates (SE), hard exudates (EX) and hemorrhages (HE) and OD are provided as a binary mask for performance evaluation of individual lesion segmentation techniques. It includes color fundus images and binary masks made of lesions. In addition to all the abnormalities, binary masks for the optic disc region are provided for all 81 images for the purpose of OD localization and detection. Furthermore, the data set provides information regarding disease severity level by grading each image in the database as diabetic retinopathy and diabetic macular edema (DME) based on international standards of clinical relevance. The medical experts graded the full set of 516 images with a variety of pathological conditions of DR and DME.

We performed image segmentation on the images present in the optic disc folder of the data set to evaluate the precision of OVIBLMM in detecting and localizing the OD in the DR and DME images. Figure 4.5 illustrates two of the best segmented clusters generated by the OVIBLMM. These images are later merged as a post processing step in order to compare it with the ground truth and measure the accuracy of the algorithm. It should be noted that due to various classification of the DR images into DR and DME, different images gave different accuracy however, we averaged the accuracies to conclude the performance of our algorithm.

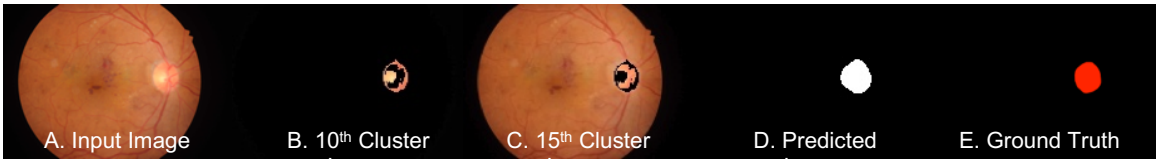

Figure 4.5: Example of best segmented optic disc (OD) images for OD detection in retinal fundus image. In the full panel, from left to right are : A. Input DR image, B. 10^{th} Cluster Image, C. 15^{th} Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image

Table 4.2 below depicts the performance of the applied OVIBLMM as compared to

[2]https://idrid.grand-challenge.org/Data/

OVGIDMM, OVIDMM and OVGMM. The algorithm performs better in terms of accuracy for all evaluation metrics.

Method	Dice	Jaccard
OVIBLMM	**99.24**	**98.5**
OVGIDMM	90.50	85.70
OVIDMM	85.54	77.76
OVGMM	79.47	68.45

Table 4.2: Evaluation metrics for optic disc detection in DR. OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy.

4.3.3 Skin Melanoma Detection

In this section, we implement OVIBLMM to effectively test its melanoma region segmentation of dermoscopic images. The algorithm was implemented on the open source data set from International Skin Imaging Collaboration (ISIC)[3] [61]. The data set contains a total of 23,906 dermoscopic images with ground truth being provided for each image.

Figure 4.6 represents the resulting two of the best segmented images along with the input image, predicted image generated by the implementation of OVIBLMM algorithm and ground truth. It is seen that the predicted image (Figure 3.4 C) is very similar to the ground truth image marked by the dermatologists (Figure 3.4 D).

[3]https://www.isic-archive.com

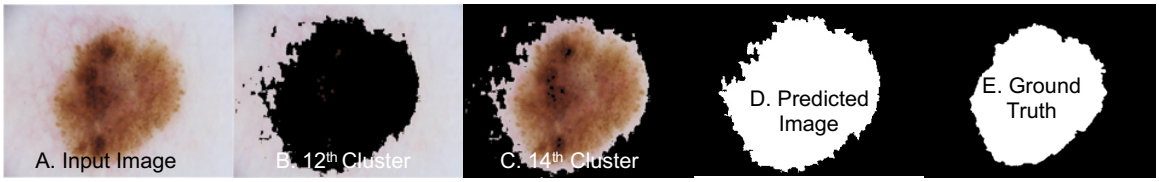

Figure 4.6: Example of best segment generated from the implementation OVIBLMM algorithm on dermoscopic images of melanoma. In the full panel, from left to right are : A. Input melanoma image from the data set, B. 12^{th} Cluster Image, C. 14^{th} Cluster Image, D. Predicted Image, E. Ground Truth Image

Table 4.3 below shows the accuracy and performance comparison of OVIBLMM as compared to OVGIDMM, OVIDMM and OVGMM. For all evaluation metrics, the algorithm proposed in this chapter is superior in terms of accuracy. Thus, the proposed algorithm is able to segment the skin melanoma accurately and therefore can be utilised and applied in research of melanoma.

Method	ARI	AMI	V-Measure	Dice	Jaccard
OVIBLMM	**93.34**	**79.86**	**85.39**	**98.07**	**96.17**
OVGIDMM	87.92	74.09	78.99	95.8	92.09
OVIDMM	74.77	61.57	66.52	90.33	81.85
OVGMM	49.33	39.77	44.64	69.48	58.23

Table 4.3: Evaluation metrics for skin melanoma detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy.

4.3.4 Colon Cancer Detection

Colorectal cancer is the third most common cause of cancer related death world wide after prostate and lung carcinoma and usually affects men and women aged over 50 years [107]. Most colon cancers initially develop a colorectal polyps, which is a small clump of cells that grows on the lining of the colon or rectum that can later become cancerous. There are two main categories of polyps, non-neoplastic and neoplastic. Non-neoplastic polyps include hyperplastic, inflammatory polyps and hamartomatous polyps. Typically, these types of polyps are not cancerous. Neoplastic polyps include adenomas and serrated types. In general the larger a polyp, the greater the risk of cancer, especially with neoplastic polyps. The identification of such polyps is carried out by MRI, which is the primary imaging modality for the diagnosis [108]. Furthermore, complete colon segmentation and detection of polyps is done by virtual computed tomography (colonoscopy or CTC) which scans the abdomen [109].

A common approach involved in colon segmentation includes the following three steps. (i) Removal of air around the body (ii) Masking of air contained in the lungs (iii) Segmentation of the colon cancer into different slices. However, the above steps do not provide the desired results in all scenarios. The above steps are also difficulties and they make the segmentation of colon more complicated, especially for implementation of an automated algorithm [110].

Human HT-29 colon cancer cells are commonly used in biology to understand the colon neoplasms and development of colorectal cancer at cellular and molecular scales. When analysing the microscopic images of such cell lines, millions of cells are present in this kind of high throughput screening (HTS). Visual classification of each of the cells into different phenotype becomes infeasible due millions of cells. Therefore, in this section, we use the publicly available human HT-29 microscopy image data set from Broad Bioimage Benchmark Collection (BBBC018v1)[4].

The set of images consists of 56 fields of view (4 for each of 14 samples). There are a total of 168 images due to three channels for the different stains applied on the cell lines. The samples were stained with Hoechst 33342, pH3 and phalloidin. Hoechst 33342 is a

[4]https://data.broadinstitute.org/bbbc/BBBC018/

DNA stain that identifies the nucleus. Phospho-histone H3 indicates mitosis. Phalloidin tags actin that is present in the cytoplasm. Each image in the data set is 512 x 512 pixels which is in DIB format 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 as classification of nucleus from the cytoplasm is an important step towards segmentation and understanding morphological abnormalities. Therefore, we test the robustness of OVIBLMM on the data set by analysing the accuracy of the algorithm on its identification of the cell morphologies.

In Figure 4.7, the representative segmentation of the actin channel from the image data set is shown. As described previously, this channel is stained by phallodin and it indicates the cytoplasmic morphology as actins are highly present in the cytoplasm. Two of the best segmented clusters generated by the algorithm are merged as a post processing step in order to compare with the ground truth. It is worth mentioning that in each image, the algorithm could identify lot of cluster of image segments, approximately upto the number of components. However, we merged the best segmented images in order to compare to the ground truth.

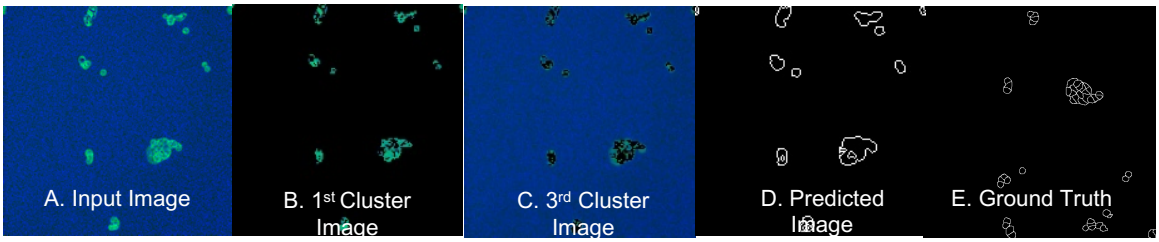

Figure 4.7: Representative best segment of human HT-29 colon cancer cells where the cytoplasm is segmented. In the full panel, from left to right are: A. Input Colon actin image, B. 1st Cluster Image, C. 3rd Cluster Image, D. Predicted Image from post processing, E. Ground Truth Image

Table 4.4 below shows the result obtained from the proposed algorithm in this chapter: OVIBLMM which is compared to OVGIDMM, OVIDMM and OVGMM. It is seen that the algorithm is superior in terms of accuracy for all evaluation metrics. This proves that the algorithm is capable of accurate image segmentation and therefore can be applied in approaches where the detection of cellular morphologies has to be automated.

Method	Dice	Jaccard
OVIBLMM	**97.29**	**94.77**
OVGIDMM	95.3	92.40
OVIDMM	92.78	90.86
OVGMM	84.32	81.51

Table 4.4: Evaluation metrics for human colon cancer detection where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Dice and Jaccard. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy.

4.3.5 Computer Aided Detection (CAD) of Malaria

In this section, we employed OVIBLMM algorithm on the malaria blood smear data set from NIH [5]. The data set has been described in the previous chapter. There are 151 patient entries for the parasitized class and the uninfected class includes 201 entries as the normal cells. In this experiments, the feature extraction concept played a very crucial role in this data set to evaluate the performance of our algorithm. The features were extracted using the BOVW, SIFT, and color histogram method. Feature extraction has the target of decreasing the subsequent computational complication and facilitating a credible and accurate recognition for unknown new data. For this experiment, the number of components M had been initialized with 2. The initial values of the hyper-parameters u, p, g, v and s for the conjugate priors are fixed to 1, q, h and t were set to 10. a and η_o the learning rate parameters are fixed to 0.5 and 64 respectively . In the Figure 3.6 and Figure 3.7 are example result images from the application of the algorithm, The images of parasitized cells are illustrated in Figure 3.6 and Figure 3.7 illustrates the images of uninfected cells detected by the algorithm.

The accuracy and performance of OVIBLMM algorithm against other algorithms are depicted in Table 4.5. Compare to other three algorithms, OVIBLMM gives greater accuracy, recall, precision and F1- Score for CAD images of malaria . In the case of malaria

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

data set, it is a cellular data set (RBC images) as opposed to tissue data set in the previous examples. In this application we show that the OVIBLMM is able to also classify data set which is on cellular scale. Therefore, these results exhibit that online variational learning is a robust method for heterogeneous biological and biomedical data set.

Method	Accuracy	Precision	Recall	F1-score
OVIBLMM	**93.5**	**96.88**	**92.76**	**92.40**
OVGIDMM	91.2	95.10	82.11	88.88
OVIDMM	90.3	87.96	79.47	87.59
OVGMM	83.80	75.26	92.71	83.08

Table 4.5: Evaluation metrics for malaria data set where the OVIBLMM was compared to OVGIDMM; OVIDMM and OVGMM. The evaluation metrics chosen were Accuracy, Precision, Recall and F1- Score. It is seen that OVIBLMM performs the best from the above algorithms, giving the highest accuracy.

Chapter 5

Conclusion

Biomedical and medical data are essential and complex data to analyze accurately and efficiently [14]. Artificial intelligence (AI) techniques have greatly improved segmentation precision due to their ability to tackle complex information. Thus, in many cases, AI has lent to early diagnosis and treatment of diseases by assisting the doctors as CAD support systems. In this work, we introduced three new statistical approaches for online variational learning framework based on three different distributions namely finite inverted Dirichlet, finite generalized inverted Dirichlet with feature selection and finite IBL mixture model to analyze multimodal images of biomedical origins. We have shown the segmentation accuracy of the online variational learning approach on medically diverse data set (tissue as well as cellular) along with its robustness to deal with different imaging modalities as input image file.

Table 5.6 summarises the data sets on which the algorithm was implemented in our study where we used different modalities of images to evaluate our models performance in each case.

However, it has to be noted that medical data comes with some challenges. The most important challenge is the limited availability of data set with ground truth. Collecting annotated cases in medical imaging is often a tough task. Furthermore, as more and more imaging modalities are being implemented, performing the annotation on new images will also be tedious and expensive. Therefore, implementing the algorithms on various available data sets and then broadening their application on data set with no ground truth is the way forward. It is therefore important to have an understanding of the correct initial parameters

Organ	Dataset Name	Dataset Size	Modality	Format
Brain	BRATS2015	189	MRI	.MHA
Lung	NIH	138	X-Ray	.PNG
Brain	LGG MRI Segmentation	110	MRI	.TIF
Retina	IDRiD	81	Retinal Fundus camera	.TIF
Skin	ISIC	23,906	Digital camera	.JPG
HT-29(Cellular)	BBC018	168	Microscopy	.DIB
RBC(Cellular)	NIH	27,558	Microscopy	.PNG

Table 5.6: Overview of the biomedical data sets on which our models have been implemented in this study. The data set is heterogeneous in nature (various organs and modalities).

to train the model such that we can transfer the algorithm to data set without ground truth and obtain an accurate segmentation. Another challenge in medical data segmentation is the heterogeneous appearance of the organs. There is a huge variance in shape, size and location of lesions or abnormalities in the images from patient to patient. While segmenting different target tissues or cells on the data set, it is of extreme importance to consider the relevant information on the image by automating the algorithm specifically for particular data set. Further to this, medical imaging has a strong implementation of 3D approaches to be able to evaluate the prognosis of the disease with precision and without invasive techniques. For such an application, the algorithm introduced in this thesis can be implemented on converting the 3D image to 2D.

The learning process in our approach is based upon variational inference in an online manner and permits closed-form solutions for the various involved model parameters. Variational learning provides good generalization capabilities, but at a significant lower computational cost since it does not need calculations of high-dimensional integrals as in MCMC methods. The approach allows analytical calculations of posterior distributions over the mixtures hidden variables, parameters, and structure [111]. Thus, we were able to determine the model parameters and the number of components simultaneously within the framework. The proposed framework of online variational learning as an extension to batch algorithm keeps not only the advantages of previous models, but also speeds up the convergence rate significantly.

In all the models we have implemented in our study, OVIBLMM outperforms the other

models of OVGIDMM, OVIDMM and OVGMM. However, each of the models has its own advantages which have been briefly discussed in chapter 2, 3 and 4 respectively. We believe this study may help researches to choose one of the algorithms to perform image segmentation on medical data and also be aware of the possible challenges and the solutions. Future work could be dedicated to integrating feature selection in the proposed model of OVI-BLMM to have higher accuracy and better performance. Another potential future work is extending the proposed framework via non-parametric Bayesian techniques.

Bibliography

[1] Menze B, Jakab A, Bauer S, Kalpathy-Cramer J, Farahaniy K, Kirby J, Burren Y, Porz N, Slotboomy J, Wiest R, Lancziy L, Gerstnery E, Webery MA, Arbel T, B Avants B, Ayache N, Buendia P, Collins L, Cordier N, Van Leemput K (2014) The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 99

[2] Lai RZ, Sakellaropoulos G, Kistler M, Bonaretti S, Pfahrer M, Niklaus R, Büchler P (2013) The virtual skeleton database: An open access repository for biomedical research and collaboration. Journal of Medical Internet Research

[3] Agrawal JP, Erickson BJ, Kahn CE (2016) Imaging Informatics: 25 Years of Progress. Yearb Med Inform Suppl 1:23–31

[4] Sohail MN, Jiadong R, Uba MM, Irshad M (2019) A comprehensive looks at data mining techniques contributing to medical data growth: A survey of researcher reviews. In: Patnaik S, Jain V (eds) Recent Developments in Intelligent Computing, Communication and Devices, Springer Singapore, Singapore, pp 21–26

[5] Ganguly D, Chakraborty S, Balitanas M, Kim Th (2010) Medical imaging: A review. In: Kim Th, Stoica A, Chang RS (eds) Security-Enriched Urban Computing and Smart Grid, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 504–516

[6] Perera CM, Chakrabarti R (2015) A review of m-health in medical imaging. Telemedicine and e-Health 21(2):132–137

[7] Lester DS, Olds JL (2001) Biomedical imaging: 2001 and beyond. The Anatomical Record: An Official Publication of the American Association of Anatomists 265(2):35–36

[8] Van Beek EJ, Hoffman EA (2008) Functional imaging: Ct and mri. Clinics in chest medicine 29(1):195–216

[9] Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics 31(4-5):198–211

[10] Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S, Freedman MT, Fryd D, Gur D, Hadjiiski L, et al. (2013) Evaluation of computer-aided detection and diagnosis systemsa. Medical physics 40(8)

[11] Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. Radiographics 37(2):505–515

[12] Guadalupe Sanchez M, Guadalupe Sánchez M, Vidal V, Verdu G, Verdú G, Mayo P, Rodenas F (2012) Medical image restoration with different types of noise. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 4382–4385

[13] Pham DL, Xu C, Prince JL (2000) Current methods in medical image segmentation. Annual Review of Biomedical Engineering 2(1):315–337

[14] Concepts DM (2006) Technique, jiawei han and micheline kamber. University of Illinois at Urbana-Champaign,

[15] Banfield JD, Raftery AE (1993) Model-based gaussian and non-gaussian clustering. Biometrics pp 803–821

[16] Bouguila N, Ziou D, Vaillancourt J (2003) Novel mixtures based on the dirichlet distribution: application to data and image classification. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer, pp 172–181

[17] Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis & Machine Intelligence (3):381–396

[18] Zhu J, Ge Z, Song Z (2017) Variational bayesian gaussian mixture regression for soft sensing key variables in non-gaussian industrial processes. IEEE Transactions on Control Systems Technology 25(3):1092–1099

[19] Bdiri T, Bouguila N (2012) Positive vectors clustering using inverted dirichlet finite mixture models. Expert Syst Appl 39(2):1869–1882

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

[21] Kalra M, Osadebey M, Bouguila N, Pedersen M, Fan W (2020) Online Variational Learning for Medical Image Data Clustering, Springer International Publishing, pp 235–269

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

[23] Tirdad P, Bouguila N, Ziou D (2015) Variational Learning of Finite Inverted Dirichlet Mixture Models and Applications, Springer International Publishing, Cham, pp 119–145

[24] Ganesalingam S (1989) Classification and mixture approaches to clustering via maximum likelihood. Journal of the Royal Statistical Society Series C (Applied Statistics) 38(3):455–466

[25] McLachlan G, Krishnan T (2007) The EM algorithm and extensions, vol 382. John Wiley & Sons

[26] Fan W, Bouguila N, Ziou D (2012) Variational learning for finite dirichlet mixture models and applications. IEEE Transactions on Neural Networks and Learning Systems 23(5):762–774

[27] Christian R, Casella G (1999) Monte carlo statistical methods (book review). Technometrics 42(4):430

[28] Gultepe E, Makrehchi M (2018) Improving clustering performance using independent component analysis and unsupervised feature learning. Human-centric Computing and Information Sciences

[29] Fan W, Bouguila N, Ziou D (2014) Variational learning of finite dirichlet mixture models using component splitting. Neurocomputing 129:3–16

[30] Zakariya SM, Ali R, Ahmad N (2010) Combining visual features of an image at different precision value of unsupervised content based image retrieval. In: 2010 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2010

[31] Fan W, Bouguila N (2012) Online variational learning of finite dirichlet mixture models. Evolving Systems 3(3):153–165

[32] Fan W, Bouguila N (2014) Online variational learning of generalized dirichlet mixture models with feature selection. Neurocomputing 126:166 – 179, recent trends in Intelligent Data Analysis Online Data Processing

[33] Constantinopoulos C, Likas A (2007) Unsupervised learning of Gaussian mixtures based on variational component splitting. IEEE Transactions on Neural Networks

[34] Williams G (2011) Descriptive and predictive analytics. In: Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Springer New York, New York, NY, pp 171–177

[35] Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 15(1):55–86

[36] Bellazzi R, Zupan B (2008) Predictive data mining in clinical medicine: Current issues and guidelines

[37] Swan M (2009) Emerging patient-driven health care models: An examination of health social networks, consumer personalized medicine and quantified self-tracking. International Journal of Environmental Research and Public Health 6:492 – 525

[38] Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A (2018) Clinical Data Mining: a Review. Yearbook of Medical Informatics

[39] Chechulin Y, Nazerian A, Rais S, Malikov K (2014) Predicting patients with high risk of becoming high-cost healthcare users in ontario(canada). Healthcare policy = Politiques de sante 9:68–79

[40] Ramezankhani A, Kabir A, Pournik O, Azizi F, Hadaegh F (2016) Classification-based data mining for identification of risk patterns associated with hypertension in middle eastern population: A 12-year longitudinal study. Medicine 95

[41] Parva E, Boostani R, Ghahramani Z, Paydar S (2017) The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literatrue. Bull Emerg Trauma

[42] Kuo IT, Chang KY, Juan DF, Hsu SJ, Chan CT, Tsou MY (2018) Time-dependent analysis of dosage delivery information for patient-controlled analgesia services. PLoS ONE

[43] Ming-Jang L, Chao-Ju C, King-Teh L, Hon-Yi S (2015) Trend Analysis and Outcome Prediction in Mechanically Ventilated Patients: A Nationwide Population-Based Study in Taiwan. PLoS One

[44] Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S (2018) Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS ONE

[45] Tiao GG, Cuttman I (1965) The inverted dirichlet distribution with applications. Journal of the American Statistical Association 60(311):793–805

[46] Xu R, Wunsch DC (2010) Clustering algorithms in biomedical research: A review

[47] xian Wang H, Bin L, bing Zhang Q, Wei S (2004) Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm. Pattern Recognition Letters 25:1799–1809

[48] Schneider A, Hommel G, Blettner M (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt international

[49] Kovalchuk SV, Funkner AA, Metsker OG, Yakovlev AN (2018) Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification. Journal of Biomedical Informatics

[50] Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: Towards better research applications and clinical care

[51] Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational Inference: A Review for Statisticians

[52] Corduneanu A, Bishop CM (2001) Variational bayesian model selection for mixture distributions. In: Artificial intelligence and Statistics, Morgan Kaufmann Waltham, MA, vol 2001, pp 27–34

[53] Lawrence N, Bishop C, Jordan M (1998) Mixture representations for inference and learning in Boltzmann machines. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence

[54] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Machine Learning 37(2):183–233

[55] Bishop CM, Lawrence N, Jaakkola T, Jordan MI (1998) Approximating posterior distributions in belief networks using mixtures. In: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10, MIT Press, Cambridge, MA, USA, NIPS '97, pp 416–422

[56] Amari SI (1998) Natural gradient works efficiently in learning. Neural Comput 10(2):251–276

[57] Hoffman M, Bach FR, Blei DM (2010) Online learning for latent dirichlet allocation. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (eds) Advances in Neural Information Processing Systems 23, Curran Associates, Inc., pp 856–864

[58] Bakas S, Kuijf HJ, Menze BH, Reyes M (2017) Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. In: Lecture Notes in Computer Science

[59] Barkhof F, Scheltens P (2002) Imaging of white matter lesions. Cerebrovascular Diseases

[60] Arroyo-Camarena S, Domínguez-Cherit J, Lammoglia-Ordiales L, Fabila-Bustos DA, Escobar-Pio A, Stolik S, Valor-Reed A, de la Rosa-Vázquez J (2016) Spectroscopic and Imaging Characteristics of Pigmented Non-Melanoma Skin Cancer and Melanoma in Patients with Skin Phototypes III and IV. Oncology and Therapy

[61] Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra NK, Kittler H, Halpern A (2017) Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). CoRR abs/1710.05006, 1710.05006

[62] Asaid R, Boyce G, Padmasekara G (2012) Use of a smartphone for monitoring dermatological lesions compared to clinical photography. Journal of Mobile Technology in Medicine 1:16–18

[63] Wu X, Marchetti MA, Marghoob AA (2015) Dermoscopy: not just for dermatologists. Melanoma Manag 2(1):63–73

[64] Sakamoto K (2012) The Pathology of Mycobacterium tuberculosis Infection

[65] Huda W, Abrahams RB (2015) Radiographic techniques, contrast, and noise in x-ray imaging. AJR Am J Roentgenol 204(2):W126–131

[66] Brady A, Ó Laoide R, McCarthy P, Mcdermott R (2012) Discrepancy and error in radiology: Concepts, causes and consequences

[67] Candemir S, Jaeger S, Palaniappan K, P Musco J, K Singh R, Xue Z, Karargyris A, Antani S, Thoma G, Mcdonald C (2014) Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Transactions on Medical Imaging 33:577–590

[68] Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Xue Z, Palaniappan K, Singh RK, Antani S, Thoma G, Wang Y, Lu P, McDonald CJ (2014) Automatic tuberculosis screening using chest radiographs. IEEE Transactions on Medical Imaging 33(2):233–245

[69] Mashrgy MA, Bdiri T, Bouguila N (2014) Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. Knowledge-Based Systems 59:182 – 195

[70] McLachlan G, Peel D (2004) Finite mixture models. John Wiley & Sons

[71] Bdiri T, Bouguila N, Ziou D (2016) Variational bayesian inference for infinite generalized inverted dirichlet mixtures with feature selection and its application to clustering. Applied Intelligence 44(3):507–525

[72] Boutemedjet S, Bouguila N, Ziou D (2009) A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(8):1429–1443

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

[74] Bishop C, Winn J (2003) Structured variational distributions in vibes. In: Proceedings Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, pp 3–6, iSBN 0-9727358-0-1

[75] Chandler D (1987) Introduction to Modern Statistical Mechanics

[76] Celeux G, Forbes F, Peyrard N (2003) Em procedures using mean field-like approximations for markov model-based image segmentation. Pattern recognition 36(1):131–144

[77] Sato MA (2001) Online model selection based on the variational bayes. Neural computation 13(7):1649–1681

[78] Kushner H, Yin GG (2003) Stochastic approximation and recursive algorithms and applications, vol 35. Springer Science & Business Media

[79] Demirhan A, Törü M, Güler I (2014) Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks. IEEE journal of biomedical and health informatics 19(4):1451–1458

[80] Roy K, Mukherjee J (2013) Image similarity measure using color histogram, color coherence vector, and sobel method. International Journal of Science and Research 2:538–543

[81] Cooper G (2000) The cell: A molecular approach. Sinauer associates

[82] Sharma K, Kaur A, Gujral S (2014) Brain tumor detection based on machine learning algorithms. International Journal of Computer Applications 103(1)

[83] Arber A, Faithfull S, Plaskota M, Lucas C, De Vries K (2010) A study of patients with a primary malignant brain tumour and their carers: symptoms and access to services. International journal of palliative nursing 16(1):24–30

[84] Merchant TE, Pollack IF, Loeffler JS (2010) Brain tumors across the age spectrum: biology, therapy, and late effects. Semin Radiat Oncol 20(1):58–66

[85] Villanueva-Meyer JE, Mabray MC, Cha S (2017) Current Clinical Brain Tumor Imaging. Neurosurgery 81(3):397–415

[86] Işın A, Direkoğlu C, Şah M (2016) Review of mri-based brain tumor image segmentation using deep learning methods. Procedia Computer Science 102:317–324

[87] Bandhyopadhyay SK, Paul TU (2013) Automatic segmentation of brain tumour from multiple images of brain mri. Int J Appl Innovat Eng Manage (IJAIEM) 2(1):240–8

[88] Meena A, Raja R (2013) Spatial fuzzy c means pet image segmentation of neurodegenerative disorder. ArXiv abs/1303.0647

[89] Cernazanu-Glavan C, Holban S (2013) Segmentation of bone structure in x-ray images using convolutional neural network. Adv Electr Comput Eng 13(1):87–94

[90] Yerpude A, Dubey S (2012) Colour image segmentation using k-medoids clustering. Int J Comput Technol Appl 3(1):152–4

[91] Ogden S, Telfer NR (2009) Skin cancer. Medicine 37(6):305–308

[92] Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, Bilek P, Braun-Falco O, Plewig G (1994) The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology 30(4):551–559

[93] Leonardi GC, Falzone L, Salemi R, Zanghi A, Spandidos DA, Mccubrey JA, Candido S, Libra M (2018) Cutaneous melanoma: From pathogenesis to therapy (Review). Int J Oncol 52(4):1071–1080

[94] Goulart JM, Quigley EA, Dusza S, Jewell ST, Alexander G, Asgari MM, Eide MJ, Fletcher SW, Geller AC, Marghoob AA, et al. (2011) Skin cancer education for primary care physicians: a systematic review of published evaluated interventions. Journal of general internal medicine 26(9):1027

[95] Das DK, Ghosh M, Pal M, Maiti AK, Chakraborty C (2013) Machine learning approach for automated screening of malaria parasite using light microscopic images. Micron 45:97–106

[96] Watanabe K, Akaho S, Omachi S, Okada M (2009) Variational bayesian mixture model on a subspace of exponential family distributions. IEEE transactions on neural networks 20(11):1783–1796

[97] Wagner RF, Metz CE, Campbell G (2007) Assessment of medical imaging systems and computer aids: a tutorial review. Academic radiology 14(6):723–748

[98] Maanicshah K, Azam M, Nguyen H, Bouguila N, Fan W (2020) Finite inverted beta-liouville mixture models with variational component splitting. In: Mixture Models and Applications, Springer, pp 209–233

[99] Hu C, Fan W, Du JX, Bouguila N (2019) A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. Neurocomputing 333:110 – 123

[100] Wang W, Lo ACY (2018) Diabetic Retinopathy: Pathophysiology and Treatments. Int J Mol Sci 19(6)

[101] Zhang X, Saaddine JB, Chou CF, Cotch MF, Cheng YJ, Geiss LS, Gregg EW, Albright AL, Klein BE, Klein R (2010) Prevalence of diabetic retinopathy in the united states, 2005-2008. Jama 304(6):649–656

[102] Ting DSW, Tan KA, Phua V, Tan GSW, Wong CW, Wong TY (2016) Biomarkers of diabetic retinopathy. Current diabetes reports 16(12):125

[103] Duh EJ, Sun JK, Stitt AW (2017) Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. JCI Insight 2(14)

[104] Stitt AW, Curtis TM, Chen M, Medina RJ, McKay GJ, Jenkins A, Gardiner TA, Lyons TJ, Hammes HP, Simo R, Lois N (2016) The progress in understanding and treatment of diabetic retinopathy. Prog Retin Eye Res 51:156–186

[105] Usman Akram M, Khan A, Iqbal K, Butt WH (2010) Retinal images: Optic disk localization and detection. In: Campilho A, Kamel M (eds) Image Analysis and Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 40–49

[106] Hatef E, Vanderver BG, Fagan P, Albert M, Alexander M (2015) Annual diabetic eye examinations in a managed care Medicaid population. Am J Manag Care 21(5):297–302

[107] Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G (2006) User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31(3):1116–1128

[108] Bert A, Dmitriev I, Agliozzo S, Pietrosemoli N, Mandelkern M, Gallo T, Regge D (2009) An automatic method for colon segmentation in ct colonography. Computerized Medical Imaging and Graphics 33(4):325–331

[109] Kekelidze M, D'Errico L, Pansini M, Tyndall A, Hohmann J (2013) Colorectal cancer: current imaging methods and future perspectives for the diagnosis, staging and therapeutic response evaluation. World J Gastroenterol 19(46):8502–8514

[110] Gayathri Devi K, Radhakrishnan R (2015) Automatic segmentation of colon in 3D CT images and removal of opacified fluid using cascade feed forward neural network. Comput Math Methods Med 2015:670739

[111] Fan W, Bouguila N (2013) Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. IEEE Transactions on Neural Networks and Learning Systems 24:1850–1862

A Appendix

A.1 Proof of equation (2.17): Variational solution of $Q(\mathcal{Z})$

For the variational solution $Q_s(\Theta_s)$, the general expression is expressed as:

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

where $const$ is an additive term representing every term that is independent of $Q_s(\Theta_s)$. Now consider the joint distribution in equation (2.10), the variational solution for $Q(\mathcal{Z})$ can be derived as follows:

$$\ln Q(\mathcal{Z}) = \alpha_{ij} \left[\ln \pi_j + \mathcal{R}_j + \sum_{l=1}^{D+1} (\alpha_{jl} - 1) \ln X_{il} \right] + const \tag{A.2}$$

where

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

and

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

Since we don't have a closed form solution for $\mathcal{R}_j$, therefore it is not possible to directly apply the variational inference. Therefore in order to provide traceable approximations, the second order Taylor's expansion is used to approximate the expected values of parameters α_j [23]. Hence, considering the logarithm form of (2.6) the equation $(A.2)$ can be written as

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

where

$$\ln \rho_{ij} = \ln \pi_j + \mathcal{R}_j + \sum_{l=1}^{D} (\alpha_{jl} - 1) \ln X_{il} \tag{A.6}$$

Since all the term without $\mathcal{Z}_{ij}$ can be added to the constant, it possible to show that

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

To find the exact formula for $Q(\mathcal{Z})$, equation (53) should be normalized and the calculation can be expressed as

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

where

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

It is noteworthy that $\sum_{j=1}^{M} r_{ij} = 1$, thus the result for $Q(\mathcal{Z})$ is

$$\big\langle \mathcal{Z}_{ij} \big\rangle = r_{ij} \tag{A.10}$$

A.2 Proof of equation $\big(2.18\big)$, $\big(2.22\big)$ and $\big(2.23\big)$

Assuming the parameters α_{jl} are independent in a mixture model with M components, we can factorize $Q(\alpha)$ as

$$Q(\alpha) = \prod_{j=1}^{M}\prod_{l=1}^{D+1} Q(\alpha_{jl}) \tag{A.11}$$

We compute the variational solution for the $\mathrm{Q}\big(\alpha_{jl}\big)$ by using the equation $\big(2.16\big)$ instead of using the gradient method. The logarithm of the variational solution $\mathrm{Q}\big(\alpha_{jl}\big)$ is given by,

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

where,

$$\mathcal{J}\big(\alpha_{jl}\big) = \left\langle \ln \frac{\Gamma\big(\alpha_{jl}+\sum_{s\neq l}^{D+1}\alpha_{js}\big)}{\Gamma\big(\alpha_{jl}\big)\prod_{s\neq l}^{D+1}\Gamma\big(\alpha_{js}\big)} \right\rangle_{\Theta\neq\alpha_{jl}} \tag{A.13}$$

Similar to what we encountered in the case of R_j the equation for $\mathcal{J}\big(\alpha_{jl}\big)$ is also intractable. We solve this problem finding the lower bound for the equation by calculating the first-order

Taylor expansion with respect to $\overline{\alpha}_{jl}$. The calculated lower bound is given by [31],

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

Substituting this equation for lower bound in equation $\big(A.12\big)$

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

This equation can be rewritten as,

$$\ln Q\big(\alpha_{jl}\big) = \ln \alpha_{jl} \big(u_{jl} + \varphi_{jl} - 1 \big) - \alpha_{jl} \big(\nu_{jl} - \vartheta_{jl} \big) + \text{const} \tag{A.16}$$

where,

$$\begin{aligned} \varphi_{jl} = & \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \overline{\alpha}_{jl} \Bigg[\psi\Bigg(\sum_{l=1}^{D+1} \overline{\alpha}_{jl} \Bigg) - \psi\big(\overline{\alpha}_{jl}\big) \\ & + \sum_{s \neq l}^{D+1} \overline{\alpha}_{js} \psi'\Bigg(\sum_{l=1}^{D+1} \overline{\alpha}_{jl} \Bigg) \big(\big\langle \ln \alpha_{js} \big\rangle - \ln \overline{\alpha}_{js} \big) \Bigg] \end{aligned} \tag{A.17}$$

$$\vartheta_{jl} = \sum_{i=1}^{N} \big\langle Z_{ij} \big\rangle \Bigg[\ln X_{il} - \ln \Bigg(1 + \sum_{l=1}^{D} X_{il} \Bigg) \Bigg] \tag{A.18}$$

Equation $\big(A.16\big)$ is the logarithmic form of a Gamma distribution. If we exponentiate both the sides, we get,

$$Q\big(\alpha_{jl}\big) \propto \alpha_{jl}^{u_{jl} + \varphi_{jl} - 1} e^{-\big(\nu_{jl} - \vartheta_{jl}\big) \alpha_{jl}} \tag{A.19}$$

This leaves us with the optimal solution for the hyper-parameters u_{jl} and ν_{jl} given by,

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

A.3 Proof of equation (2.27)

We calculate the mixing coefficients value π by maximizing the lower bound w.r.t to π. It is essential to include Lagrangian term in the lower bound because of the constraint $\sum_{j=1}^{M} \pi_j = 1$. Then, solving for the derivative w.r.t π_j and setting the result to zero, we have [31]

$$\begin{aligned}\frac{\partial \mathcal{L}(Q)}{\partial \pi_j} &= \frac{\partial \mathcal{L}(Q)}{\partial \pi_j} \sum_{i=1}^{N} \sum_{j=1}^{M} r_{ij} \ln \pi_j + \lambda \left(\sum_{j=1}^{M} \pi_j - 1 \right) \\ &= \sum_{i=1}^{N} r_{ij}(1/\pi_j) + \lambda = 0 \end{aligned} \tag{A.21}$$

$$\Rightarrow \sum_{i=1}^{N} r_{ij} = -\lambda \pi_j \tag{A.22}$$

By taking the sum of both sides of equation (A.22) over j, we can obtain $\lambda = -N$. Then substituting the value of λ equation (A.21), we can obtain

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{A.23}$$