Novel Mixture Allocation Models for Topic Learning

Kamal Maanicshah

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

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

March 2023

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared:

By: **Kamal Maanicshah**
Entitled: **Novel Mixture Allocation Models for Topic Learning**

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

Doctor of Philosophy of "Information and Systems Engineering"

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

Signed by the final examining committee:

Dr. Belkacem Chikhaoui	———	External Examiner
Dr. Joonhee Lee	———	External Examiner
Dr. Roch Glitho	———	Internal Examiner
Dr. Mohsen Ghafouri	———	Internal Examiner
Dr. Nizar Bouguila	———	Supervisor
Dr. Manar Amayri	———	Supervisor

Approved by ———
Dr. Abdessamad Ben Hamza, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

March, 2023 ———
Dr. Mourad Debbabi
Dean of Faculty of Engineering and Computer Science

Abstract

Novel Mixture Allocation Models for Topic Learning

Kamal Maanicshah
Concordia University, 2023

Unsupervised learning has been an interesting area of research in recent years. Novel algorithms are being built on the basis of unsupervised learning methodologies to solve many real world problems. Topic modelling is one such fascinating methodology that identifies patterns as topics within data. Introduction of latent Dirichlet Allocation (LDA) has bolstered research on topic modelling approaches with modifications specific to the application. However, the basic assumption of a Dirichlet prior in LDA for topic proportions, might not be applicable in certain real world scenarios.

Hence, in this thesis we explore the use of generalized Dirichlet (GD) and Beta-Liouville (BL) as alternative priors for topic proportions. In addition, we assume a mixture of distributions over topic proportions which provides better fit to the data. In order to accommodate application of the resulting models to real-time streaming data, we also provide an online learning solution for the models. A supervised version of the learning framework is also provided and is shown to be advantageous when labelled data are available.

There is a slight chance that the topics thus derived may not be that accurate. In order to alleviate this problem, we integrate an interactive approach which uses inputs from the user to improve the quality of identified topics. We have also tweaked our models to be applied for interesting applications such as parallel topics extraction from multilingual texts and content based recommendation systems proving the adaptability of our proposed models. In the case of multilingual topic extraction, we use global topic proportions sampled from a Dirichlet process (DP) to tackle the problem and in the case of recommendation systems, we use the co-occurrences of words to our advantage.

For inference, we use a variational approach which makes computation of variational solutions easier. The applications we validated our models with, show the efficiency of proposed models compared to other state of the art alternatives for the same tasks.

Acknowledgments

I would like to express my deepest gratitude to my supervisor Prof. Nizar Bouguila, who has been patient with me and supported me throughout my work. I owe the progress in my thesis to his constant motivation and faith in me through the ups and downs. I will always be grateful for his relentless support and guidance.

I express my profound gratitude to my co-supervisor Dr. Manar Amayri for her contributions.

I am grateful for the support provided by MITACS (accelerate program) and giving me an opportunity to work with brilliant minds such as Andree at Ciena. It was wonderful to work with some awesome colleagues and learn from the opportunity.

I was also fortunate to have wonderful colleagues who have kept me motivated throughout. I am always grateful for the fun times and insightful discussions I had with Narges, Muhammed, Hussein, Hafsa, Fatma, Ornela, Rim, Pantea, Eddy, Ons and Oumayma to name a few. They form the best part of my PhD memories. A special thanks goes to Narges and Eddy who have also contributed their knowledge for the completion of this thesis.

Apart from my colleagues in the lab I am glad that I had such wonderful housemates and friends who understood me and stood by me during troubling situations. Last but not least, I am deeply grateful to my parents who encouraged me to work hard and sent loads of love virtually which kept pushing me forward each day.

Contents

List of Figures

List of Tables

List of Acronyms

LDA: Latent Dirichlet Allocation

GD: Generalized Dirichlet

BL: Beta-Liouville

DP: Dirichlet Process

LSA: Latent Semantic Analysis

SVD: Singular Value Decomposition

TF-IDF: Term Frequency - Inverse Document Frequency

PLSA: Probabilistic Latent Semantic Analysis

MCMC: Markov Chain Monte Carlo

ANN: Artificial Neural Network

LGDA: Latent generalized Dirichlet allocation

LBLA: Latent Beta-Liouville allocation

LGDMA: Latent generalized Dirichlet mixture allocation

LBLMA: Latent Beta-Liouville mixture allocation

SVM: Support Vector Machine

MLE: Maximum likelihood estimation

BoW: Bag of Words

KL: Kullback-Leibler

sLGDMA: Supervised Latent generalized Dirichlet mixture allocation

sLBLMA: Supervised Latent Beta-Liouville mixture allocation

KNN: K-Nearest Neighbors

RF: Random Forest

DNA: Deoxyribonucleic Acid

RNA: Ribonucleic Acid

HoG: Histogram of Gradients

SIFT: Scale Invariant Feature Transform

iLGDMA: Intereactive latent generalized Dirichlet mixture allocation

iLBLMA: Interactive latent Beta-Liouville mixture allocation

Bi-LGDMA: Biterm latent generalized Dirichlet mixture allocation

Bi-LBLMA: Biterm latent Beta-Liouville mixture allocation

API: Application Programming Interface

DP-LGDMA: DP latent generalized Dirichlet mixture allocation

DP-LBLMA: DP latent Beta-Liouville mixture allocation

NTM: Neural Topic Models

VAE: Variational Auto Encoders

GANs: Generative Adversarial Networks

Chapter 1

Introduction

1.1 Introduction and Related Work

In recent years, owing to improved data collection techniques and the introduction of sensors and devices in multiple domains, the need to analyze and understand a plethora of accumulated data is an important task. A number of data analysis techniques based on machine learning have been proposed over the years to analyze and understand the data. These techniques help us identify inherent patterns and unique structure of the data. Topic modelling is one such technique used for identifying patterns from documents in the form of topics. This thesis introduces novel statistical models for topic modelling and elaborates how they could be easily tailored to specific applications.

Topic modelling techniques were basically introduced for document retrieval tasks [1, 2]. However, the concept has been expanded to other domains like image analysis, genome pattern recognition, etc. The underlying principle of topic models is to identify topics within a document which are represented by the distribution of words in the vocabulary. Each document is considered to have a combination of these topics. This helps us in unsupervised content retrieval tasks [3]. This identified topics can also be used for classification tasks [4] and as feature extraction methods [5].

Topic modelling algorithms have evolved since the introduction of latent semantic analysis (LSA) [1]. Latent Dirichlet allocation (LDA) proposed in [2] was a major turn around in topic modelling research. The basic LDA model has been modified since then to suit various applications in multiple

domains. Further features were also integrated with the standard model to make it more efficient. For example, in [6], the authors propose spatial LDA, which incorporates spatial information between patches within an image region which helps in object detection. The idea revolves around the fact that parts of an object close together might be topics belonging to the same object. It is a semi-supervised approach designed for object detection. To integrate time related properties to the topics learned, the authors in [7] propose to use Hawkes process to model the frequency of texts along with the topics learned from LDA to identify fake re-tweeters. In [8], a model is presented, which takes into account the correlation between the topics. The authors in [9], propose an online version of LDA to handle streaming real-time data. One of the main attributes of LDA is that it learns the latent patterns in an unsupervised manner. The work in [10] proposes a supervised method without compromising the unsupervised learning part. Many more models have been proposed recently to solve prevalent machine learning problems.

However, there are a few challenges that are to be considered while designing a topic model:

- The assumption of Dirichlet prior in LDA though proven to be efficient implies a negative covariance matrix by definition and might not provide a good fit for the document topic distribution for certain datasets. In these cases a preferable option would be to use an alternative prior with similar properties that overcome the drawbacks of Dirichlet distribution.

- Though the model might identify words that dominate a particular topic, there is a chance that the topic could be adulterated with a few words that might not belong to that topic.

- The choice of estimation method for the parameters also place a huge role in the efficiency of the model. In general most of the topic modelling algorithms use approximation methods such as variational inference or pure Bayesian method like Gibbs sampling.

1.2 Contributions

- Introduction of novel topic models for topic extraction based on mixtures of distributions.

- A modified variational inference method, that eases the calculation of variational solutions for the proposed models. In addition we also show how these models can be modified easily to online and supervised use cases.

- Application of the model to multi domain applications like image classification, genome classification and text classification.

- We take into account the presence of biterms in text data to build models inculcating this information. These models are evaluated based on their recommendation capabilities for recommendation systems.

- We show how the models can be used with Dirichlet process to be applied for multilingual texts.

1.3 Thesis Overview

- Chapter 2: This chapter, serves as a preamble to our work detailing the various concepts used to build the models. A brief introduction to LDA, mixture models, etc. is given.

- Chapter 3: In this chapter we propose novel models namely latent generalized Dirichlet mixture allocation (LGDMA) and latent generalized Beta-Liouville mixture allocation (LBLMA) models. We also detail how these models can be converted to adopt to supervised and online learning scenarios. We evaluate the models with standard text mining tasks, in addition to genome and image classification.

- Chapter 4: In this chapter we explore the option to make our model interactive with user inputs to extract better topics. Generally, the topics identified by LBLMA and LGDMA models might sometime consist of a few words which are irrelevant to that particular topic. An interactive algorithm would help us to use insights from the user to improve the quality of topics learned. Hence, we propose interactive LGDMA (iLGDMA) and interactive LBLMA (iLBLMA) models which use user inputs to extract better topics.

- Chapter 5: There might be cases where the subsequent words in a document might have an impact on each other in terms of the topic

they belong to. This use case is our concern in this chapter and we focus on creating a model which would help us to learn these related topics. This gives rise to the Bi-term mixture allocation models. We use the models to build recommendation systems for anime and movie suggestions.

- Chapter 6: Here we introduce a non parametric approach to extract multilingal topics from parallel corpora using Dirichlet process mixture allocation models. This helps us in indexing similar topics across multiple languages which helps in multilingual document retrieval. In addition this model avoids the need for a model selection method. We test our models against multilingual datasets on Ted talks subtitles which are on different subjects.

- Chapter 7: This chapter summarizes the results obtained with our work and explains future potential research work.

1.4 Publications and Submissions

Five manuscripts have been built out of the content of this thesis. Among them, two are submitted to journals, two have been accepted in conferences and another one is under review in a conference. The details of the manuscripts follow:

- Chapter 3:
  - Kamal Maanicshah, Manar Amayri, Nizar Bougila, "Novel Mixture Allocation Models for Topic Learning" is submitted to the journal "Computational Intelligence"

- Chapter 4:
  - Kamal Maanicshah, Manar Amayri, Nizar Bougila, "Interactive Generalized Dirichlet Mixture Allocation" is published in "joint IAPR international workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)"
  - Kamal Maanicshah, Manar Amayri, Nizar Bougila, "Improving Topic Quality with Interactive Beta-Liouville Mixture Allocation

Model" is published in "IEEE Symposium Series On Computational Intelligence (SSCI)"

- Chapter 5:
  - Kamal Maanicshah, Manar Amayri, Nizar Bougila, "Novel Topic Models for Content based Recommender Systems" is accepted at "International Conference on Enterprise Information Systems (ICEIS)"
- Chapter 6:
  - Kamal Maanicshah, Narges Manouchehri, Manar Amayri, Nizar Bougila, "Novel Topic Models for Parallel Topics Extraction from Multilingual Text" is submitted to "International Journal of Computational Intelligence Systems"

Chapter 2

Background and Preliminary Concepts

When it comes to machine learning tasks, we can categorize them into three broad categories namely, supervised, unsupervised and reinforcement learning based on their methodology [11]. Reinforcement learning is a method where an agent learns by exploration to react to its environment [12]. Supervised learning can be used when we have labelled data and the task involves identifying patterns within them [10]. However, labelled data are not widely available for a number of tasks and hence we are in need of unsupervised learning approaches like topic modelling, mixture models, etc, which can learn patterns from data irrespective of the labels [13, 14]. This helps us to cluster the data which in turn can be used to label in bulk or directly for application specific pattern recognition tasks like bag of topics creation [15], document indexing [16–18], software moudule categorization [19], image categorization [20], spam filtering [21], etc.

This chapter details the evolution of topic models starting with latent semantic analysis in section 2.1 followed by probabilistic latent semantic analysis and LDA in sections 2.2 and 2.3 respectively. The basic concepts required for understanding our models like mixture models, parameters estimation and choice of distribution are explained in sections 2.4, 2.5 and 2.6, followed by other challenges faced in section 2.8.

2.1 Latent Semantic Analysis

Topic modelling techniques caters to a set of models that are capable of recognizing patterns within data in the form of topics. LSA proposed in [1] can be considered as the origin for research in topic modelling approaches. This simple model uses the basic logic that words with similar themes appear more frequently together. The model uses singular value decomposition (SVD) on the term frequency - inverse document frequency (TF-IDF) matrix which quantizes the occurrences of words in a vocabulary. This helps to reduce the dimension of the feature space. If we assume that we have a document term matrix given by M, then SVD decomposes this matrix into three components as follows:

$$M = U\Sigma V^T \tag{2.1}$$

where, U is the document topic matrix for K topics which is user defined, Σ is the covariance matrix for the topics and V is the term topic matrix respectively. The matrix U can be effectively used for document retrieval purposes. The main drawback of this model is that we do not have an interpretable version of the topics which could be of use for other decision making tasks like labelling, document understanding, etc.

2.2 Probabilistic Latent Semantic Analysis

Later the same concept of identifying relevant features in documents was rebuilt to follow a probabilistic approach [22]. The model follows a generative technique, assuming that the probability can be generated from a set of topics. Fig. 2.1 shows the graphical representation of the model. Here, for

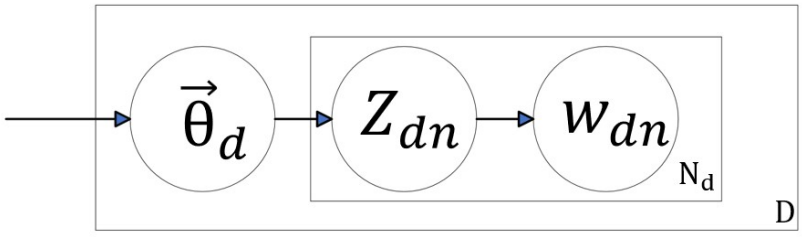

Figure 2.1: Graphical representation of PLSA

a set of D documents, w_{dn} represents the n^{th} word among N_d words in the

document d, Z_{dn} indicates the affinity of the word to topics and $\vec{\theta}_d$ represent the probability of document d to be generated from the topics. We can write the probability of documents to be generated as,

$$p(d, w_{dn}) = p(\vec{\theta}_d) \sum_Z p(z_{dn} \mid w_{dn}) p(z_{dn} \mid \vec{\theta}_d) \tag{2.2}$$

It is notable that this representation looks like a mixture model where $p(\vec{\theta}_d)$ can be considered as a vector of mixing proportions. However, in this case, though we can get a good representation of the topics in the training data, it is not possible to assign proportions to a newly seen document since each representation of the document is like a fixed point in the dataset.

2.3 Latent Dirichlet Allocation

LDA [2] adds a Bayesian flavor to PLSA by using Dirichlet priors for the topic word and document topic proportions. The topic word proportion is the probability of each word in the vocabulary to belong to a particular topic and the document topic proportions refer to the probability of that document to belong to each of the topics. Let us consider a corpus containing D documents which is represented as a bag of words (BoW) model. The BoW model represents each document with a vector indicating the frequency of each word in the vocabulary; $\vec{w}_d = \{w_{dn}\}$, where $n = 1, 2, ..., N_d$ is the n^{th} word among the N_d words in the document d. It is to be noted that each word w_{dn} can also be represented as a vector of V dimensions where V is the size of the vocabulary. For the v^{th} word in the vocabulary, $w_{dnv} = 1$ when word $w_{dn} = v$ and 0 elsewhere. The belongingness of a word to a particular topic k is denoted by a latent variable $\mathcal{Z} = \{\vec{z}_d\} = \{\vec{z}_{dn}\}$ with $z_{dnk} = 1$ when the word belongs to topic k. The probability that a word $w_{dnv} = 1$ when $z_{dnk} = 1$ is assumed to be drawn from a multinomial distribution with parameter $\vec{\beta} = \{\vec{\beta}_k\} = \{\beta_{kv}\}$. The document topic proportions of LDA, $\theta_{dk} = p(z_{dnk} = 1)$ is defined by a Dirichlet distribution with parameters $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_K)$ as mentioned in Eq. 2.10. With these assumptions, the likelihood of a set of documents $W = \{\vec{w}_d\}$ over the topic proportions and

topic assignments is given by the equation,

$$p(W \mid \vec{\sigma}, \vec{\tau}, \vec{\beta}) = \prod_{d=1}^{D} \int \Big[p(\vec{\theta}_d \mid \vec{\sigma}) \times \prod_{n=1}^{N_d} \sum_{k=1}^{K} p(w_{dnv} = 1 \mid z_{dnk} = 1, \vec{\beta}) p(z_{dnk} = 1 \mid \vec{\theta}_d) \Big] d\vec{\theta}_d \tag{2.3}$$

where, $p(\vec{\theta}_d \mid \vec{\sigma})$ is the Dirichlet prior with parameter $\vec{\sigma}$. The graphical model of LDA is shown in Fig. 2.2. Based on this setup, when a new document is looked at, a document topic proportion can still be drawn from the Dirichlet distribution which overcomes the drawbacks of PLSA.

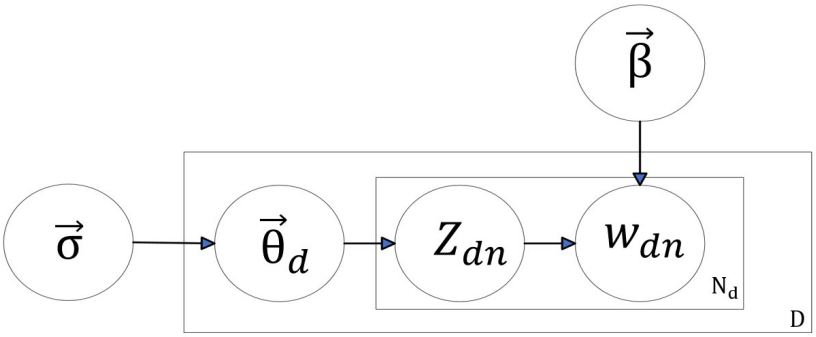

Figure 2.2: Graphical representation of LDA

2.4 Mixture Models

Similar to LDA, mixture models correspond to an important section of model-based approaches for unsupervised learning of patterns within data [23, 24]. We consider the data to be derived from a mixture of distributions and estimate the parameters of these distributions, which help in further tasks like classification, prediction, image segmentation, image retrieval, etc. [25–27]. Though Gaussian mixture models [28] are widely used, there are others which use non Gaussian assumption to provide a better fit to the data [29,30]. There has been recent studies which combine both the concept of mixture models

and LDA as explored in [31, 32] where the document topic proportions are assumed to be sampled from a mixture of Dirichlet distributions as opposed to one distribution in the case of LDA. This makes the model flexible to provide a better fit to the data.

Basically, we consider that the data is generated by a weighted sum of a number of distributions. This assumption makes mixture models useful for clustering tasks as each different characteristic or pattern within the data might be represented by a mixture component. Consider a dataset with D data points represented by $\mathcal{X} = (\vec{X}_1, \vec{X}_2, ..., \vec{X}_D)$. Each data point is a vector of N dimensions. Assuming, there are L components within this data, we can write the equation for the mixture model as,

$$p(\vec{X}_i \mid \vec{\pi}, \Theta) = \sum_{l=1}^{L} \pi_l p(\vec{X}_i \mid \theta_l) \tag{2.4}$$

where, $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_L)$, $\Theta = \{\theta_1, \theta_2, ..., \theta_L\}$ and $p(\vec{X}_i \mid \theta_l)$ for the i^{th} document is the probability distribution from which the data is assumed to be sampled from and can be any distribution such as Gaussian [33], Dirichlet [34], Beta [35], etc. The parameter $\vec{\pi}$ represents the mixing coefficients and strictly follows the constraints $0 \leq \pi_l \leq 1$ and $\sum_{l=1}^{L} \pi_l = 1$. Let's introduce an indicator matrix $\mathcal{Z} = (\vec{z}_1, \vec{z}_2, ..., \vec{z}_D)$ which shows the cluster membership of each of the data points. Each vector $\vec{z}_i$ is a L-dimensional vector $\vec{z}_i = (z_{i1}, z_{i2}, ..., z_{iL})$ with $z_{il} = 1$ if data point i belongs to the cluster component l and 0 otherwise. With this assumption, the conditional distribution of $\mathcal{Z}$ given the mixing weights $\vec{\pi}$ can be sampled from a multinomial given by,

$$p(\mathcal{Z} \mid \vec{\pi}) = \prod_{i=1}^{D} \prod_{l=1}^{L} \pi_l^{z_{il}} \tag{2.5}$$

Introducing this equation above in Eq. 2.4, we can write the general complete likelihood of a mixture model as,

$$p(\mathcal{X} \mid \mathcal{Z}, \Theta) = \prod_{i=1}^{D} \prod_{l=1}^{L} \Big(\pi_l p(\vec{X}_i \mid \theta_l)\Big)^{z_{il}} \tag{2.6}$$

2.5 Estimation of Parameters

Once the design of the model is finalized, the next challenging task is to estimate the model parameters. Several approaches have been proposed to

estimate the parameters of mixture models such as maximum likelihood [36], Bayesian inference [37–39], variational Bayes [29], expectation propagation [40], etc. Some of the methods are explained below:

2.5.1 Maximum Likelihood Estimation

When a set of data points are defined by a model, maximum likelihood estimation (MLE) involves determining the optimal set of parameters values that maximize the likelihood [41]. The likelihood of a model is the joint probability of all the observed data points. We can easily find the maxima of the likelihood function by finding the derivative of the log likelihood and equating it to zero. However, depending on initialization values, this method might result in identifying parameter values which are actually saddle points which may not result in good models to fit well the data.

2.5.2 Bayesian Approach

Pure Bayesian statistical inference methods comprises of sampling algorithms such as Markov Chain Monte Carlo (MCMC) sampling [42, 43]. MCMC involves sampling from the target distribution until an approximate value of the true posterior is found. Gibbs sampling is one such method which samples new data points based on the previous samples. Though Gibbs sampling provides a more accurate solution for the parameters, it is hard for one to evaluate the convergence of these pure Bayesian methods [44]. Variational inference on the other hand provides an approximate solution for the parameters instead of trying to find the true solution. Gibbs sampling and variational inference [2, 45] are the most commonly used methods for estimating posterior distribution in topic modelling approaches. The only drawback of variational algorithms is that they suffer from some bias due to initialization. In our models we use variational inference for estimating the posterior owing to its simplicity.

2.5.3 Variational Inference

Let us assume a Bayesian framework defined by a set of N data points denoted by $D = \{d_1, d_2, ..., d_N\}$ with latent variables and parameters defined by $Y = \{y_1, y_2, ..., y_N\}$. We can find the joint distribution based on our probabilistic model given by $p(D, Y)$. The objective is to find the posterior

distribution $p(D \mid Y)$. The idea of variational inference is to approximate this posterior to a variational distribution. Let's say $p(Q)$ is the variational distribution. This variational distribution can be approximated to be the true posterior by minimizing the distance between them. We can do this by calculating the Kullback-Leibler (KL) divergence between the two distributions, given by,

$$KL(Q \mid\mid P) = -\int Q\big(Z\big) \ln\left(\frac{p\big(D \mid Z\big)}{Q\big(Z\big)}\right) dZ \tag{2.7}$$

Simplifying this equation we have,

$$KL(Q \mid\mid P) = \ln p(D) - \mathcal{L}(Q) \tag{2.8}$$

where,

$$\mathcal{L}(Q) = \int Q\big(Z\big) \ln\left(\frac{p\big(D, Z\big)}{Q\big(Z\big)}\right) dZ \tag{2.9}$$

From these equations, we can see that maximizing the lower bound given by $\mathcal{L}(Q)$ minimizes the KL divergence between the true posterior P and variational distribution Q. Based on this idea, an approximate solution for the true posterior can be found.

In general the variational approach followed for topic models is as shown in [2]. Here the variational lower bound is used as a surrogate for the marginal likelihood. By setting the derivative of the lowerbound to zero the maxima for each of the parameters can be calculated. However, using this method resulted in intractable solutions for our model. Hence we adopted the variational method followed by the authors in [29] which made the derivation of solutions easier. In this case, we work on deriving a variational solution which are in the form of the prior distribution. The update functions of the two can be found by equating the parameters. The method is clearly explained in Appendix A.

2.6 Choice of Distributions

It is also important to consider the choice of distributions for the priors to achieve better performance with topic models. A few of them that concerns our models are described in this section.

2.6.1 Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of Beta distribution. A Dirichlet distribution is defined by,

$$Dir(\vec{\theta} \mid \sigma) = \frac{\Gamma(\sum_{k=1}^{K} \sigma_k)}{\prod_{k=1}^{K} \Gamma(\sigma_k)} \prod_{k=1}^{K} \theta_k^{\sigma_k - 1} \tag{2.10}$$

where, $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_K)$ represent the parameters of a random variable $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_K)$ with K dimensions. The Dirichlet distribution has a negative covariance structure. However, there is a possibility that the covariance might be positive in which case Dirichlet distribution might not give a good fit to the data. There has been some research on finding an alternative for Dirichlet distribution as prior for the document topic proportions. For example, authors in [46–49] use Poisson point process, generalized Dirichlet (GD) and Beta-Liouville (BL) as efficient alternatives for the Dirichlet prior.

2.6.2 Generalized Dirichlet Distribution

GD is one such distribution which has been shown to provide a better fit for those random variables with general covariance [50–53]. Consider a random variable $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_K)$, where K is the dimension of the vector. $\vec{\theta}$ follows the constraints $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k < 1$. The probability density function of $\vec{\theta_k}$ following a GD distribution is given by,

$$GD(\vec{\theta} \mid \vec{\sigma}, \vec{\tau}) = p(\vec{\theta} \mid \vec{\sigma}, \vec{\tau}) = \prod_{k=1}^{K} \frac{\Gamma(\sigma_k + \tau_k)}{\Gamma(\sigma_k)\Gamma(\tau_k)} \theta_k^{\sigma_k - 1} \left(1 - \sum_{j=1}^{k} \theta_j\right)^{\gamma_k} \tag{2.11}$$

where, $\vec{\tau} = (\tau_1, \tau_2, ..., \tau_K)$, $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_K)$ with $\gamma_k = \tau_k - \tau_{k+1} - \sigma_{k+1}$ for $k = 1, 2, ..., K-1$ and $\gamma_k = \sigma_k - 1$ for $k = K$.

2.6.3 Beta-Liouville Distribution

Though the GD seems to be an effective choice, twice the number of parameters as compared to the Dirichlet is to be estimated. This led to use the BL distribution, in some works, as an alternative since it also allows a positive covariance matrix [54–56]. Similar to the previous definition, let

$\vec{\theta} = (\theta_1, \theta_2, ..., \theta_K)$ be the random variable with K dimensions with the same constraints $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k < 1$. The BL distribution is defined as,

$$BL(\vec{\theta} \mid \vec{\mu}, \sigma, \tau) = \prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} \mu_k)}{\Gamma(\mu_k)} \frac{\Gamma(\sigma + \tau)}{\Gamma(\sigma)\Gamma(\tau)} \theta_k^{\mu_k - 1} \times \Big[\sum_{k=1}^{K} \theta_k\Big]^{\sigma - \sum_{k=1}^{K} \mu_k} \Big[1 - \sum_{k=1}^{K} \theta_k\Big]^{\tau - 1} \tag{2.12}$$

where, $\vec{\mu} = (\mu_1, \mu_2, ..., \mu_K)$, σ and τ are parameters of the Beta-Liouville distribution. These distributions can serve as effective priors to be used for topic modelling tasks [48].

2.6.4 Dirichlet Process

A Dirichlet process (DP) can be thought of as an extension of the Dirichlet distribution, where each sample can be considered as a distribution in itself. This can be of help for infinite cases where the model complexity is not known. There are a few realizations of DP that are widely used [57, 58]. In this thesis we use the stick breaking definition to build our model as it makes inference easier. According to this definition, when a stick of length l is broken, we assume the length of the broken section to be sampled from a Beta distribution, $\beta_c \sim Beta(1, \alpha)$. Here α is the parameter of Beta distribution and c represents the c^{th} broken piece from the stick. If $\{\pi_c\}_{c=1}^{\infty}$ is the total number of pieces the stick is broken into, then the length of each broken stick can be defined as,

$$\pi_c = \beta_c \prod_{j=1}^{c-1} \Big(1 - \beta_j\Big) \tag{2.13}$$

This stick-breaking representation of DP could be used as prior for the mixing weights in mixture models [59].

2.7 Evaluation Methods for Topic Models

Evaluation of topic models is a challenging task since it is an unsupervised approach. However, there are some methods usually followed to check the quality of extracted topics. The basic approach would be to list the top

words in each topic to see if they represent that topic. However, following eye balling approaches like this would make it hard for us to compare results from different models. Quantitative evaluation of topic models can be divided into two groups: extrinsic and intrinsic methods.

2.7.1 Extrinsic Methods

Extrinsic methods does not directly measure the quality of the extracted topics. Instead, we evaluate the quality based on the task at hand. For example, if we are using the extracted topics for a classification task, the best quality topics might consequently result in good classification accuracy. In Chapters 3 and 5 we evaluate our models based on classification accuracy and recommendation relevancy which are good examples of this kind.

2.7.2 Intrinsic Methods

Evaluation methods such as perplexity, topic coherence, etc. come under the umbrella of intrinsic evaluation. Perplexity is one of the methods used for topic evaluation. According to this method, we split the data set into training and testing sets and find the perplexity of each document in the test set, based on the formula,

$$Perplexity(\vec{w}_{d\_test}) = \exp\left\{-\frac{\sum_{d=1}^{D} \ln p(\vec{w}_d)}{\sum_{d=1}^{D} N_d}\right\} \tag{2.14}$$

Here, $\vec{w}_{d\_test}$ is the word vector for a test document, $\vec{w}_d$ is the word vector of d^{th} document in the training set containing D documents and N_d is the number of words in the d^{th} document. However, perplexity is not considered to be an accurate depiction of topic quality because it considers only the occurrence of the particular word within the topic. In addition, the method also requires us to split the data into train and test sets which might affect the quality as well.

In general coherence measures which include the co-occurrences of words within the topic tend to reflect the quality of topics more accurately. UMass coherence score [60], is one such method used to calculate the relevancy of words within the topic. For each of the topics, the coherence score is

calculated by using the formula,

$$score_{UMass}(k) = \sum_{i=2}^{M_k} \sum_{j=1}^{M_k-1} \log \frac{p(w_i, w_j) + 1}{p(w_i)} \tag{2.15}$$

where M_k is the set of words in a topic with w_i and w_j being the i^{th} and j^{th} words in the topic. This metric is used for evaluating our models in Chapters 4 and 6.

2.8 Other Challenges

The main advantage of topic models in general is that it can be easily adopted to handle challenges specific to the data. The novel structure of the models introduced in this work also proposed its own challenges which need to be addressed. This thesis shows how our models can be easily altered to solve these challenges. Here are some of the challenges addressed in this thesis:

Adulterated Topics

It is a pressing issue in topic models that some of the topics identified may contain words that does not belong to that topic. It would be an interesting improvement if we are able to interactively provide input to the model regarding the correctness of the learned topics [61]. However, integrating interactive learning poses some challenges as well. It is important to keep in mind that in the case of large datasets it would be impossible for the users to check all the discovered topics. This calls for a proper mechanism to identify which topics we need to show to the user on priority for modification. It is also our primary desire to preserve the unsupervised topic extraction capability of our proposed models. These challenges are clearly addressed and detailed in Chapter 4 of the thesis.

Structure of Text

Presence of words that co-occur together is a known property in texts. In certain cases the effects of these bigrams can be ignored as they might not hurt the efficiency of the model considerably. In certain data however, the number of bigrams might be higher and might be important to consider.

Chapter 5 of this thesis focuses on this problem by introducing a biterm model which considers the presence of bigrams in the documents.

Model Selection

Experiments in Chapters 3, 4 and 5 show that, due to the introduction of mixture model based design in our models, though it improves the topic learning capabilities, we have to identify the optimal number of mixture components. In general, it is a general norm in topic modelling approaches to identify the number of topics for optimal representation. Pinning down the number of components becomes an additional burden for our models. Hence, in Chapter 6, we propose a DP based model which can take care of the model selection problem.

Chapter 3

Mixture Allocation Models

Over the years, extensive research has been conducted to learn patterns from data. This helps in various decision making tasks in many industries such as manufacturing, healthcare, etc. Recent developments in artificial neural networks (ANN) based models have made them very useful for various tasks. However, it is a well known fact that deep learning models are more like a black box that work well for the task but lack interpretability which is quintessential for tasks that need more reasoning [62]. On the other hand, classic machine learning models like LDA, mixture models, etc. have good interpretability even though they might achieve a slightly lower accuracy when compared to deep learning models.

Owing to the proven efficiency of using GD and BL priors in topic models, we introduce a topic model based on latent generalized Dirichlet allocation (LGDA) [63] and latent Beta-Liouville allocation (LBLA) [47] combining it with mixture models to enhance the support of respective topics giving rise to latent generalized Dirichlet mixture allocation (LGDMA) and latent Beta-Liouville mixture allocation (LBLMA) models, respectively.

In order to improve the modelling capabilities, we use variational inference method for estimating the parameters. Additionally, we also introduce an online variational approach to cater to specific applications involving streaming data. We evaluate our models based on its performance on applications related to text classification, image categorization and genome sequence classification using a supervised approach where the labels are used as an observed variable within the model [10]. This addition helps us to learn topics pertaining to each class simultaneously for classification task rather

than relying on external classifiers like Naive Bayes, support vector machine (SVM), etc. which learns from the intermediate feature space derived from the LGDMA or LBLMA model. There are a lot of applications which we can use to evaluate our model in addition to text processing. In this chapter we assess our models based on three applications from different fields of research. We will have one application on image classification, one on genome sequence classification and another on text classification.

The rest of the chapter is organized as follows: The model is described in detail in section 3.1, and the parameters estimation by variational inference method is explained in section 3.2. In section 3.3, we show how the inference method has to be modified to incorporate online learning for streaming data. The supervised LGDMA model is described in section 3.4. The experimental results are discussed in section 3.5.

3.1 Model Description

The basic idea here is to create novel models where the single GD and BL priors in LGDA and LBLA are replaced by mixtures of GD and BL distributions respectively. We first explain latent generalized Dirichlet mixture allocation (LGDMA) model in subsection 3.1.1. The latent Beta-Liouville mixture allocation (LBLMA) model follows in 3.1.2 with brief modifications avoiding redundancies.

3.1.1 Latent Generalized Dirichlet Mixture Allocation

The latent generalized Dirichlet mixture allocation model (LGDMA) follows the same generative process as in LDA and other similar models with slight variations [31]. We reiterate the variables and their roles once again as used in the rest of the thesis. Let us consider a corpus of D documents, with each document d represented as a word vector $\vec{w}_d = (w_{d1}, w_{d2}, ..., w_{dN_d})$. Here N_d is the number of words in the document and each word can be represented as a V dimensional one-hot encoded vector with $w_{dnv} = 1$ where the word $w_{dn} = v$ from the vocabulary V and 0 elsewhere. As in the previous cases, we have a D dimensional topic assignment matrix $\vec{Z}_d = (\vec{z}_{d1}, \vec{z}_{d2}, ..., \vec{z}_{dN_d})$ with $\vec{z}_{dn}$ being a K dimensional one-hot encoded vector with $z_{dnk} = 1$ when word z_{dn} belongs to the topic k out of K topics. The probability that each of these words belongs to a topic k is given by the multinomial with parameters

$\vec{\beta}_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{kV})$ for each topic k among the K topics. As opposed to a single GD distribution in the case of LGDA, in our case, we define the topic proportions $\vec{\theta}_d$ by a mixture of L GD distributions with parameters $\vec{\sigma} = \{\vec{\sigma}_l\} = \{\sigma_{lk}\}$ and $\vec{\tau} = \{\vec{\tau}_l\} = \{\tau_{lk}\}$. We also have another indicator matrix pertaining to the mixture model which is denoted by, $\mathcal{Y} = (\vec{y}_1, \vec{y}_2, ..., \vec{y}_D)$ where $\vec{y}_d$ is L dimensional with $y_{dl} = 1$ when the document d belongs to cluster l. $\mathcal{Y}$ is in turn governed by a multinomial distribution with parameters $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_L)$ which is the mixing coefficient and follows the same constraints mentioned in subsection 2.4. With this setup, the generative process in the case of LGDMA is as follows:

- For each word vector $\vec{w}_d$ in the corpus:
  - Draw component l of the mixture $y_d = l \backsim Multinomial(\vec{\pi})$
  - Draw topic proportions $\vec{\theta}_d \mid y_d = l$ from a mixture of L generalized Dirichlet distributions
  - For each word n of the N_d words in document $\vec{w}_d$
    - Draw topic $z_{dn} = k \backsim Multinomial(\vec{\theta}_d)$
    - Draw word $w_{dn} = v \mid z_{dn} = k \backsim Multinomial(\vec{\beta}_{z_{dn}})$

Based on these assumptions, we can write down the marginal likelihood of a simple LGDMA model for a dataset W with D documents as:

$$p(W \mid \vec{\pi}, \vec{\sigma}, \vec{\tau}, \vec{\beta}) = \prod_{d=1}^{D} \int \Bigg[\Big(\sum_{y_d} p(\vec{\theta}_d \mid y_d, \vec{\sigma}, \vec{\tau}) p(y_d \mid \vec{\pi}) \Big) \times \prod_{n=1}^{N_d} \sum_{z_{dn}} p(w_{dn} \mid z_{dn}, \vec{\beta}) p(z_{dn} \mid \vec{\theta}_d) \Bigg] d\vec{\theta}_d \tag{3.1}$$

From Eqs. 2.11 and 2.6, $p(\vec{\theta}_d \mid \vec{y}_d, \vec{\sigma}, \vec{\tau})$ will take the form of a mixture of generalized Dirichlet distribution with L components as shown in the following equation:

$$\begin{aligned} p(\vec{\theta}_d \mid \vec{y}_d, \vec{\sigma}, \vec{\tau}) &= \prod_{l=1}^{L} \prod_{k=1}^{K} \Big(p(\theta_{dk} \mid \sigma_{lk}, \tau_{lk}) \Big)^{y_{dl}} \\ &= \prod_{l=1}^{L} \Bigg[\prod_{k=1}^{K} \frac{\Gamma(\tau_{lk} + \sigma_{lk})}{\Gamma(\tau_{lk})\Gamma(\sigma_{lk})} \theta_{dk}^{\sigma_{lk}-1} \Bigg(1 - \sum_{j=1}^{k} \theta_{dj} \Bigg)^{\gamma_{lk}} \Bigg]^{y_{dl}} \end{aligned} \tag{3.2}$$

The rest of the terms in Eq. 3.1 is given by multinomial distribution as follows:

$$p(y_d \mid \vec{\pi}) = \prod_{l=1}^{L} \pi_l^{y_{dl}} \tag{3.3}$$

$$p(w_{dn} \mid z_{dn}, \vec{\beta}) = \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \beta_{kv}^{w_{dnv}} \Big)^{z_{dnk}} \tag{3.4}$$

$$p(z_{dn} \mid \vec{\theta}_d) = \prod_{k=1}^{K} \theta_{dk}^{z_{dnk}} \tag{3.5}$$

In order to improve the parameters estimation, statisticians usually use a conjugate prior over the unknown parameters when it comes to Bayesian statistics [64]. Following in those steps, we introduce Gamma prior to the parameters $\vec{\sigma}$ and $\vec{\tau}$ since the conjugate prior of GD is non-tractable while estimating parameters with variational inference. So, the prior distributions are now given by,

$$p(\sigma_{lk}) = \mathcal{G}(\sigma_{lk} \mid \upsilon_{lk}, \nu_{lk}) = \frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})} \sigma_{lk}^{\upsilon_{lk}-1} e^{-\nu_{lk}\sigma_{lk}} \tag{3.6}$$

$$p(\tau_{lk}) = \mathcal{G}(\tau_{lk} \mid s_{lk}, t_{lk}) = \frac{t_{lk}^{s_{lk}}}{\Gamma(s_{lk})} \tau_{lk}^{s_{lk}-1} e^{-t_{lk}\tau_{lk}} \tag{3.7}$$

where, $\mathcal{G}(\cdot)$ indicates Gamma distribution. [2] mentions a process called smoothing which is used to eliminate the problem of sparsity. This is done by assuming a Dirichlet prior over the parameter $\vec{\beta}$ as,

$$p(\vec{\beta}_k \mid \vec{\lambda}_k) = \frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda_{kv})} \prod_{v=1}^{V} \beta_{kv}^{\lambda_{kv}-1} \tag{3.8}$$

For the purpose of simplifying the inference, we assume a variational distribution over $\vec{\theta}_d$ given by the equation,

$$p(\vec{\theta}_d \mid \vec{g}_d, \vec{h}_d) = \prod_{k=1}^{K} \frac{\Gamma(g_{dk} + h_{dk})}{\Gamma(g_{dk})\Gamma(h_{lk})} \theta_{dk}^{g_{dk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj} \right)^{\zeta_{dk}} \tag{3.9}$$

where, $\zeta_{dk} = h_{dk} - g_{d(k-1)} - h_{d(k-1)}$ while $k \leq K-1$ and $\zeta_{dk} = h_{dk} - 1$ when $k = K$. Figure 3.1 shows the graphical representation of the model obtained

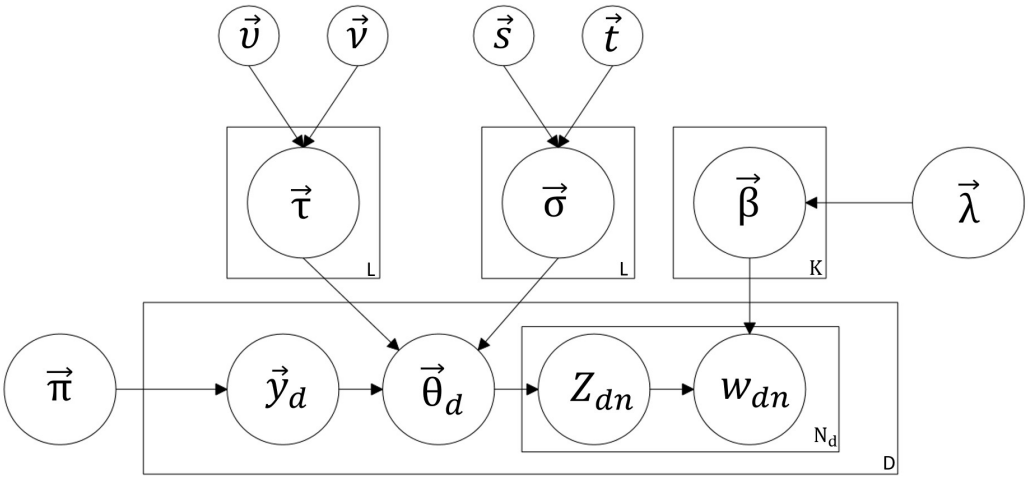

Figure 3.1: Graphical representation of LGDMA

based on these assumptions. Based on all the knowledge we have, we can write the joint distribution of the posterior as,

$$p(W,\Theta) = p(W \mid \mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\sigma}, \vec{\tau}, \vec{y}) \tag{3.10}$$

$$= p(\vec{W} \mid \mathcal{Z}, \vec{\beta}) p(\vec{z} \mid \vec{\theta}) p(\vec{\theta} \mid \vec{\sigma}, \vec{\tau}, \vec{y}) p(\vec{y} \mid \vec{\pi}) p(\vec{\theta} \mid \vec{g}, \vec{h}) p(\vec{\beta} \mid \vec{\lambda}) \times p(\vec{\sigma} \mid \vec{\upsilon}, \vec{\nu}) p(\vec{\tau} \mid \vec{s}, \vec{t}) \tag{3.11}$$

$$\begin{aligned} p(W,\Theta) = & \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \beta_{kv}^{w_{dnv}} \Big)^{z_{dnk}} \times \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{z_{dnk}} \\ & \times \prod_{d=1}^{D} \prod_{l=1}^{L} \left[\prod_{k=1}^{K} \frac{\Gamma(\tau_{lk} + \sigma_{lk})}{\Gamma(\tau_{lk})\Gamma(\sigma_{lk})} \theta_{dk}^{\sigma_{lk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj} \right)^{\gamma_{lk}} \right]^{y_{dl}} \\ & \times \prod_{d=1}^{D} \prod_{l=1}^{L} \pi_l^{y_{dl}} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \frac{\Gamma(g_{dk} + h_{dk})}{\Gamma(g_{dk})\Gamma(h_{lk})} \theta_{dk}^{g_{dk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj} \right)^{\zeta_{dk}} \\ & \times \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda_{kv})} \beta_{kv}^{\lambda_{kv}-1} \times \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})} \sigma_{lk}^{\upsilon_{lk}-1} e^{-\nu_{lk}\sigma_{lk}} \\ & \times \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{t_{lk}^{s_{lk}}}{\Gamma(s_{lk})} \tau_{lk}^{s_{lk}-1} e^{-t_{lk}\tau_{lk}} \end{aligned} \tag{3.12}$$

Given $\Theta = \{\mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\sigma}, \vec{\tau}, \vec{y}\}$ which generally represents all the parameters in the model.

3.1.2 Latent Beta-Liouville Mixture Allocation

The LBLMA model can be constructed from the same definitions considered for LGDMA model. The only difference being the prior for the topic proportions defined in Eq. 3.2 and its parameters. We replace the GD prior with the BL prior changing the set of equations to,

$$\begin{aligned} p(\vec{\theta}_d \mid \vec{y}_d, \vec{\mu}, \vec{\sigma}, \vec{\tau}) &= \prod_{l=1}^{L}\prod_{k=1}^{K}\Big(p(\theta_{dk} \mid \mu_{lk}, \sigma_l, \tau_l)\Big)^{y_{dl}} \\ &= \prod_{l=1}^{L}\prod_{k=1}^{K}\left[\frac{\Gamma(\sum_{k=1}^{K}\mu_{lk})}{\prod_{k=1}^{K}\Gamma(\mu_{lk})}\frac{\Gamma(\sigma_l+\tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)}\theta_{dk}^{\mu_{lk}-1}\right. \\ &\quad \left.\times\Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{\sigma_l-\sum_{k=1}^{K}\mu_{lk}}\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{\tau_l-1}\right]^{y_{dl}} \end{aligned} \tag{3.13}$$

This means $\vec{\theta}_d$ is assumed to be a random vector following a Beta-Liouville distribution with parameters $(\mu_{l1}, \mu_{l2}, ..., \mu_{lK}, \sigma_l, \tau_l)$. Continuing with this assumption, we can write the Gamma priors for the parameters as $p(\mu_{lk}) = \mathcal{G}(\mu_{lk} \mid \upsilon_{lk}, \nu_{lk}); p(\sigma_l) = \mathcal{G}(\sigma_l \mid s_l, t_l)$ and $p(\tau_l) = \mathcal{G}(\tau_l \mid \Omega_l, \Lambda_l)$ respectively since they have the same properties as in the case of GD. Due to this changes obviously, the variational distribution in Eq. 3.9 will be replaced by,

$$\begin{aligned} p(\vec{\theta}_d \mid \vec{f}_d, g_d, h_d) &= \prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}f_{dk})}{\prod_{k=1}^{K}\Gamma(f_{dk})}\frac{\Gamma(g_d+h_d)}{\Gamma(g_d)\Gamma(h_d)}\theta_{dk}^{f_{dk}-1} \\ &\quad\times\Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{g_d-\sum_{k=1}^{K}f_{dk}}\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{h_d-1} \end{aligned} \tag{3.14}$$

With these changes, we can construct the joint distribution of the posterior assuming a BL prior for the topic proportions as,

$$
\begin{aligned}
p(W,\Theta) = & \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K}\Big(\prod_{v=1}^{V}\beta_{kv}^{w_{dnv}}\Big)^{z_{dnk}} \times \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K}\theta_{dk}^{z_{dnk}} \\
& \times \prod_{d=1}^{D}\prod_{l=1}^{L}\Bigg[\prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}\mu_{lk})}{\prod_{k=1}^{K}\Gamma(\mu_{lk})}\frac{\Gamma(\sigma_l+\tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)}\theta_{dk}^{\mu_{lk}-1} \\
& \times \Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{\sigma_l-\sum_{k=1}^{K}\mu_{lk}}\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{\tau_l-1}\Bigg]^{y_{dl}} \times \prod_{d=1}^{D}\prod_{l=1}^{L}\pi_l^{y_{dl}} \\
& \times \prod_{d=1}^{D}\prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}f_{dk})}{\prod_{k=1}^{K}\Gamma(f_{dk})}\frac{\Gamma(g_d+h_d)}{\Gamma(g_d)\Gamma(h_d)}\theta_{dk}^{f_{dk}-1}\Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{g_d-\sum_{k=1}^{K}f_{dk}} \\
& \times \Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{h_d-1} \times \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma(\sum_{v=1}^{V}\lambda_{kv})}{\prod_{v=1}^{V}\Gamma(\lambda_{kv})}\beta_{kv}^{\lambda_{kv}-1} \\
& \times \prod_{l=1}^{L}\prod_{k=1}^{K}\frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})}\mu_{lk}^{\upsilon_{lk}-1}e^{-\nu_{lk}\mu_{lk}} \times \prod_{l=1}^{L}\frac{t_l^{s_l}}{\Gamma(s_l)}\sigma_l^{s_l-1}e^{-t_l\sigma_l} \\
& \times \prod_{l=1}^{L}\frac{\Lambda_l^{\Omega_l}}{\Gamma(\Omega_l)}\tau_l^{\Omega_l-1}e^{-\Lambda_l\tau_l} \qquad (3.15)
\end{aligned}
$$

Here $\Theta = \{\mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\mu}, \vec{\sigma}, \vec{\tau}, \mathcal{Y}\}$ indicates the set of parameters required for the model. The graphical representation of the model is shown in Fig. 3.2.

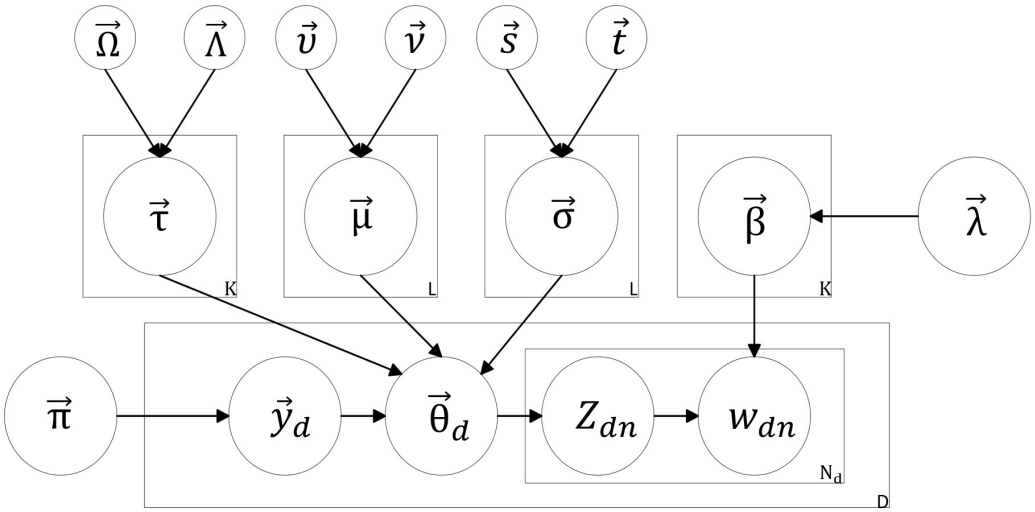

Figure 3.2: Graphical representation of LBLMA

3.2 Variational Inference

The parameter estimation method used in this work follows [64], which is slightly different from [2] that is usually used for topic models based on LDA. We choose variational inference instead of pure Bayesian methods like Gibbs sampling [45] since these algorithms might take a very long time to converge though they might give a better estimate of the parameters. The variational approach establishes a distribution $Q(\Theta)$ which is assumed to be an approximation of $p(W \mid \Theta)$ which is the posterior distribution we desire to calculate. Hence the approach conquers the shortcomings of pure Bayesian approaches by approximating the posterior rather than calculating it. According to our approach we find the similarity between the posterior and the variational distributions by using Kullback-Leibler (KL) divergence. The KL divergence between two distributions is 0 when the two distributions are similar. The KL divergence between $Q(\Theta)$ and $p(W \mid \Theta)$ is given by,

$$KL(Q \mid\mid P) = -\int Q(\Theta) \ln \left(\frac{p(W \mid \Theta)}{Q(\Theta)} \right) d\Theta \tag{3.16}$$

Simplifying this equation will lead to,

$$KL(Q \mid\mid P) = \ln p(W) - \mathcal{L}(Q) \tag{3.17}$$

where,

$$\mathcal{L}(Q) = \int Q\big(\Theta\big) \ln \left(\frac{p\big(W, \Theta\big)}{Q\big(\Theta\big)} \right) d\Theta \tag{3.18}$$

By definition of these equations, maximizing the lower bound $\mathcal{L}(Q)$ results in bringing down the KL divergence close to 0. Since the true posterior is intractable, we introduce mean-field theory [65] considering the parameters to be independent and identically distributed. Based on this idea, we can write the distribution of variational parameters as a product of individual parameters as $Q(\Theta) = \prod_{j=1}^{J} \Theta_j$ provided J is the total number of parameters. We find the optimal solution for each of the parameters by the following equation,

$$Q_j\big(\Theta_j\big) = \frac{\exp\big\langle \ln p\big(W, \Theta\big)\big\rangle_{\neq j}}{\int \exp\big\langle \ln\ p\big(W, \Theta\big)\big\rangle_{\neq j} d\Theta} \tag{3.19}$$

According to this equation, we can see that, the optimal solution for parameter Θ_j is found by calculating the expectations with respect to all the parameters other than Θ_j. Hence this process requires a suitable initialization during the start of the algorithm and then the variational solutions of each parameter are updated continuously in each iteration. This maximizes the lower bound and at convergence we find the optimal solution for all the parameters of our model. The optimal variational solutions for our LGDMA and LBLMA models are presented in the subsequent sections. A detailed explanation of how to derive the solutions is explained in Appendix A.

3.2.1 Variational solutions for LGDMA

Calculating the variational solutions for Eq. 3.12 yields the following equations:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D} \prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{k=1}^{K} \phi_{dnk}^{z_{dnk}} \tag{3.20}$$

$$Q(\vec{\sigma}) = \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{{\nu_{lk}^*}^{\upsilon_{lk}^*}}{\Gamma(\upsilon_{lk}^*)} \sigma_{lk}^{\upsilon_{lk}^* - 1} e^{-\nu_{lk}^* \sigma_{lk}} \tag{3.21}$$

$$Q(\vec{\tau}) = \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{{t_{lk}^*}^{s_{lk}^*}}{\Gamma(s_{lk}^*)} \tau_{lk}^{s_{lk}^* - 1} e^{-t_{lk}^* \tau_{lk}} \tag{3.22}$$

$$Q(\vec{\beta}) = \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma(\sum_{v=1}^{V}\lambda_{kv}^{*})}{\prod_{v=1}^{V}\Gamma(\lambda_{kv}^{*})}\beta_{kv}^{\lambda_{kv}^{*}-1} \tag{3.23}$$

$$Q(\vec{\theta}) = \prod_{d=1}^{D}\prod_{k=1}^{K}\frac{\Gamma(g_{dk}^{*}+h_{dk}^{*})}{\Gamma(g_{dk}^{*})\Gamma(h_{lk}^{*}))}\theta_{dk}^{g_{dk}^{*}-1}\left(1-\sum_{j=1}^{k}\theta_{dj}\right)^{\zeta_{dk}^{*}} \tag{3.24}$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L}\rho_{dl}}, \phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K}\delta_{dnk}}, \pi_{l} = \frac{1}{D}\sum_{d=1}^{D}r_{dl} \tag{3.25}$$

$$\rho_{dl} = \exp\left\{\ln\pi_{l} + \mathcal{R}_{l} + \sum_{k=1}^{K}(\sigma_{lk}-1)\big\langle\ln\theta_{dk}\big\rangle + \gamma_{lk}\Big\langle 1-\sum_{j=1}^{k}\theta_{dj}\Big\rangle\right\} \tag{3.26}$$

$$\delta_{dnk} = \exp(\big\langle\ln\beta_{kv}\big\rangle + \big\langle\ln\theta_{dk}\big\rangle) \tag{3.27}$$

Here, $\mathcal{R}$ is the taylor series approximations of $\big\langle\ln\frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)}\big\rangle$ and is given by,

$$\begin{aligned}\mathcal{R} =& \ln\frac{\Gamma(\overline{\sigma}+\overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\sigma})\big](\big\langle\ln\sigma\big\rangle - \ln\overline{\sigma}) \\ &+ \overline{\tau}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\tau})\big](\big\langle\ln\tau\big\rangle - \ln\overline{\tau}) \\ &+ 0.5\overline{\sigma}^{2}\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\sigma})\big]\big\langle(\ln\sigma - \ln\overline{\sigma})^{2}\big\rangle \\ &+ 0.5\overline{\tau}^{2}\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\tau})\big]\big\langle(\ln\tau - \ln\overline{\tau})^{2}\big\rangle \\ &+ \overline{\sigma}\,\overline{\tau}\,\Psi'(\overline{\sigma}+\overline{\tau})(\big\langle\ln\sigma\big\rangle - \ln\overline{\sigma})(\big\langle\ln\tau\big\rangle - \ln\overline{\tau})\end{aligned} \tag{3.28}$$

$$\begin{aligned}\upsilon_{lk}^{*} =& \upsilon_{lk} + \sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Big[\Psi\big(\overline{\sigma}_{lk}+\overline{\tau}_{lk}\big) - \Psi\big(\overline{\sigma}_{lk}\big) \\ &+ \overline{\tau}_{lk}\Psi'\big(\overline{\sigma}_{lk}+\overline{\tau}_{lk}\big)\big(\big\langle\ln\tau_{lk}\big\rangle - \ln\overline{\tau}_{lk}\big)\Big]\overline{\sigma}_{lk}\end{aligned} \tag{3.29}$$

$$\begin{aligned}s_{lk}^{*} =& s_{lk} + \sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Big[\Psi\big(\overline{\tau}_{lk}+\overline{\sigma}_{lk}\big) - \Psi\big(\overline{\tau}_{lk}\big) \\ &+ \overline{\sigma}_{lk}\Psi'\big(\overline{\tau}_{lk}+\overline{\sigma}_{lk}\big)\big(\big\langle\ln\sigma_{lk}\big\rangle - \ln\overline{\sigma}_{lk}\big)\Big]\overline{\tau}_{lk}\end{aligned} \tag{3.30}$$

$$\nu^*_{lk} = \nu_{lk} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \big\langle \ln \theta_{dk} \big\rangle \tag{3.31}$$

$$t^*_{lk} = t_{lk} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \left\langle \ln \Big[1 - \sum_{j=1}^{K} \theta_{dj} \Big] \right\rangle \tag{3.32}$$

$$g^*_{dk} = g_{dk} + \sum_{n=1}^{N_d} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \sigma_{lk} \tag{3.33}$$

$$h^*_{dk} = h_{dk} + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \tau_{lk} + \sum_{kk=k+1}^{K} \phi_{dn(kk)} \tag{3.34}$$

$$\lambda^*_{kv} = \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} \phi_{dnk} w_{dnv} \tag{3.35}$$

$$\pi_l = \frac{1}{D} \sum_{d=1}^{D} r_{dl} \tag{3.36}$$

In the above equations, $\big\langle \cdot \big\rangle$ indicates expectation of the variable and $\bar{(\cdot)}$ is the mean of the variable. The values of these expectations [66] and mean are given by,

$$\big\langle \ln \theta_{dk} \big\rangle = \sum_{j=1}^{k} \big(\Psi(g_{dk}) - \Psi(g_{dk} + h_{dk}) \big) \tag{3.37}$$

$$\Big\langle 1 - \sum_{j=1}^{k} \theta_{dj} \Big\rangle = \sum_{j=1}^{k} \big(\Psi(h_{dk}) - \Psi(g_{dk} + h_{dk}) \big) \tag{3.38}$$

$$\overline{\sigma}_{lk} = \frac{\upsilon^*_{lk}}{\nu^*_{lk}}, \big\langle \ln \sigma_{lk} \big\rangle = \Psi\big(\upsilon^*_{lk}\big) - \ln \nu^*_{lk} \tag{3.39}$$

$$\big\langle \big(\ln \sigma_{lk} - \ln \overline{\sigma}_{lk} \big)^2 \big\rangle = \big[\Psi\big(\upsilon^*_{lk}\big) - \ln \upsilon^*_{lk} \big]^2 + \Psi'\big(\upsilon^*_{lk}\big) \tag{3.40}$$

$$\overline{\tau}_{lk} = \frac{s^*_{lk}}{t^*_{lk}}, \big\langle \ln \tau_{lk} \big\rangle = \Psi\big(s^*_{lk}\big) - \ln t^*_{lk} \tag{3.41}$$

$$\left\langle \left(\ln \tau_{lk} - \ln \overline{\tau}_{lk} \right)^2 \right\rangle = \left[\Psi\left(s^*_{lk}\right) - \ln s^*_{lk} \right]^2 + \Psi'\left(s^*_{lk}\right) \tag{3.42}$$

$$\left\langle z_{dnk} \right\rangle = \phi_{dnk}, \left\langle y_{dl} \right\rangle = r_{dl}, \left\langle \ln \beta_{kv} \right\rangle = \Psi(\lambda_{kv}) - \Psi(\sum_{f=1}^{V} \lambda_{kf}) \tag{3.43}$$

$\Psi(\cdot)$ and $\Psi(\cdot)'$ in the above equations indicate the digamma and trigamma functions respectively. To find the optimal solution, our algorithm calculates equations 3.20 - 3.24 iteratively until there is no considerable change in the lower bound estimates.

3.2.2 Variational solutions for LBLMA

The variational solutions for Eq. 3.15 is more or less the same as in the previous section, except that some definitions of variables are different in addition to the obvious change in $Q(\vec{\theta})$. The variational solutions are:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D} \prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{k=1}^{K} \phi_{dnk}^{z_{dnk}} \tag{3.44}$$

$$Q(\vec{\mu}) = \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{{\nu^*_{lk}}^{\upsilon^*_{lk}}}{\Gamma(\upsilon^*_{lk})} \mu_{lk}^{\upsilon^*_{lk}-1} e^{-\nu^*_{lk}\mu_{lk}} \tag{3.45}$$

$$Q(\sigma_l) = \prod_{l=1}^{L} \frac{{t^*_l}^{s^*_l}}{\Gamma(s^*_l)} \sigma_l^{s^*_l-1} e^{-t^*_l\sigma_l} \tag{3.46}$$

$$Q(\tau_l) = \prod_{l=1}^{L} \frac{{\Lambda^*_l}^{\Omega^*_l}}{\Gamma(\Omega^*_l)} \tau_l^{\Omega^*_l-1} e^{-\Lambda^*_l\tau_l} \tag{3.47}$$

$$Q(\vec{\beta}) = \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma(\sum_{v=1}^{V} \lambda^*_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda^*_{kv})} \beta_{kv}^{\lambda^*_{kv}-1} \tag{3.48}$$

$$\begin{aligned} Q(\vec{\theta}) = & \prod_{d=1}^{D} \prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} f^*_{dk})}{\Gamma(f^*_{dk})} \frac{\Gamma(g^*_d + h^*_d)}{\Gamma(g^*_d)\Gamma(h^*_d)} \theta_{dk}^{f^*_{dk}-1} \\ & \times \Big[\sum_{k=1}^{K} \theta_{dk} \Big]^{g^*_d - \sum_{k=1}^{K} f^*_{dk}} \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big]^{h^*_d - 1} \end{aligned} \tag{3.49}$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K} \delta_{dnk}}, \pi_l = \frac{1}{D}\sum_{d=1}^{D} r_{dl} \tag{3.50}$$

$$\begin{aligned}\rho_{dl} = \exp\Bigg\{ & \ln \pi_l + \mathcal{R}_l + \mathcal{S}_l + (\mu_{lk} - 1)\big\langle \ln \theta_{dk} \big\rangle + \Big(\sigma_l - \sum_{k=1}^{K} \mu_{lk}\Big)\Big\langle \ln \big[\sum_{k=1}^{K} \theta_{dk}\big] \Big\rangle \\ & + (\tau_l - 1)\Big\langle \ln \Big[1 - \sum_{k=1}^{K} \theta_{dk}\Big] \Big\rangle \Bigg\}\end{aligned} \tag{3.51}$$

Due to intractability, we use taylor series expansions for $\big\langle \frac{\Gamma(\sum_{k=1}^{K} \sigma_{lk})}{\Gamma(\sigma_{lk})} \big\rangle$ and $\big\langle \ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \big\rangle$ denoted by $\mathcal{R}$ and $\mathcal{S}$ respectively. The approximations are given as,

$$\begin{aligned}\mathcal{R}_l = & \ln \frac{\Gamma(\sum_{k=1}^{K} \mu_{lk})}{\prod_{k=1}^{K} \Gamma(\mu_{lk})} + \sum_{k=1}^{K} \overline{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk})\Big]\big[\big\langle \ln \mu_{lk} \big\rangle - \ln \overline{\mu}_{lk}\big] \\ & + \frac{1}{2}\sum_{k=1}^{K} \overline{\mu}_{lk}^2\Big[\Psi'\big(\sum_{k=1}^{K} \overline{\mu}_{lk}\big) - \Psi'(\overline{\mu}_{lk})\Big] - \big\langle (\ln \mu_{lk} - \ln \overline{\mu}_{lk})^2 \big\rangle \\ & + \frac{1}{2}\sum_{a=1}^{K}\sum_{b=1, a\neq b}^{K} \overline{\mu}_{la}\overline{\mu}_{lb}\Big[\Psi'\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big)\big(\big\langle \ln \mu_{la} \big\rangle - \ln \overline{\mu}_{la}\big)\big(\big\langle \ln \mu_{lb} \big\rangle - \ln \overline{\mu}_{lb}\big)\Big]\end{aligned} \tag{3.52}$$

$$\begin{aligned}\mathcal{S} = & \ln \frac{\Gamma(\overline{\sigma} + \overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\big[\Psi(\overline{\sigma} + \overline{\tau}) - \Psi(\overline{\sigma})\big](\big\langle \ln \sigma \big\rangle - \ln \overline{\sigma}) \\ & + \overline{\tau}\big[\Psi(\overline{\sigma} + \overline{\tau}) - \Psi(\overline{\tau})\big](\big\langle \ln \tau \big\rangle - \ln \overline{\tau}) \\ & + 0.5\overline{\sigma}^2\big[\Psi'(\overline{\sigma} + \overline{\tau}) - \Psi'(\overline{\sigma})\big]\big\langle (\ln \sigma - \ln \overline{\sigma})^2 \big\rangle \\ & + 0.5\overline{\tau}^2\big[\Psi'(\overline{\sigma} + \overline{\tau}) - \Psi'(\overline{\tau})\big]\big\langle (\ln \tau - \ln \overline{\tau})^2 \big\rangle \\ & + \overline{\sigma}\,\overline{\tau}\,\Psi'(\overline{\sigma} + \overline{\tau})(\big\langle \ln \sigma \big\rangle - \ln \overline{\sigma})(\big\langle \ln \tau \big\rangle - \ln \overline{\tau})\end{aligned} \tag{3.53}$$

$$\begin{aligned}\upsilon_{lk}^{*} = & \upsilon_{lk} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \overline{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk}) \\ & + \Psi\Big(\sum_{k=1}^{K}\Big)\sum_{a\neq k}^{K}\big(\big\langle \ln \mu_{la} \big\rangle - \ln \overline{\mu}_{la}\big)\overline{\mu}_{la}\Big]\end{aligned} \tag{3.54}$$

$$\nu_{lk}^{*} = \nu_{lk} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\big\langle \ln \theta_{dk} \big\rangle - \Big\langle \ln \sum_{k=1}^{K} \theta_{dk} \Big\rangle \Big] \tag{3.55}$$

$$\begin{aligned} s_{l}^{*} =& s_{l} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\Psi\big(\overline{\sigma}_{l} + \overline{\tau}_{l}\big) - \Psi\big(\overline{\sigma}_{l}\big) \\ & + \overline{\tau}_{l} \Psi'\big(\overline{\sigma}_{l} + \overline{\tau}_{l}\big) \big(\big\langle \ln \tau_{l} \big\rangle - \ln \overline{\tau}_{l} \big) \Big] \overline{\sigma}_{l} \end{aligned} \tag{3.56}$$

$$t_{l}^{*} = t_{l} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big\langle \ln \Big[\sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{3.57}$$

$$\begin{aligned} \Omega_{l}^{*} =& \Omega_{lk} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\Psi\big(\overline{\tau}_{l} + \overline{\sigma}_{l}\big) - \Psi\big(\overline{\tau}_{l}\big) \\ & + \overline{\sigma}_{l} \Psi'\big(\overline{\tau}_{l} + \overline{\sigma}_{l}\big) \big(\big\langle \ln \sigma_{l} \big\rangle - \ln \overline{\sigma}_{l} \big) \Big] \overline{\tau}_{l} \end{aligned} \tag{3.58}$$

$$\Lambda_{l}^{*} = \Lambda_{l} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big\langle \ln \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{3.59}$$

$$f_{dk}^{*} = f_{dk} + \sum_{n=1}^{N_d} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \mu_{lk} \tag{3.60}$$

$$g_{d}^{*} = g_{d} + \sum_{n=1}^{N_d} \sum_{k=1}^{K} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \sigma_{l} \tag{3.61}$$

$$h_{d}^{*} = h_{d} + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \tau_{l} \tag{3.62}$$

The expectations in these equations are defined with respect to BL distribution as follows:

$$\left\langle \ln \theta_{dk} \right\rangle = \Psi(f_{dk}) - \Psi\Big(\sum_{k=1}^{K} f_{dk}\Big) + \Psi(g_d) - \Psi(g_d + h_d) \tag{3.63}$$

$$\Big\langle \sum_{k=1}^{k} \theta_{dk} \Big\rangle = \sum_{k=1}^{k} \big(\Psi(g_d) - \Psi(g_d + h_d)\big) \tag{3.64}$$

$$\Big\langle 1 - \sum_{k=1}^{k} \theta_{dk} \Big\rangle = \sum_{k=1}^{k} \big(\Psi(h_d) - \Psi(g_d + h_d)\big) \tag{3.65}$$

$$\overline{\sigma}_{lk} = \frac{\upsilon^*_{lk}}{\nu^*_{lk}}, \left\langle \ln \sigma_{lk} \right\rangle = \Psi\big(\upsilon^*_{lk}\big) - \ln \nu^*_{lk} \tag{3.66}$$

$$\left\langle \big(\ln \sigma_{lk} - \ln \overline{\sigma}_{lk}\big)^2 \right\rangle = \big[\Psi\big(\upsilon^*_{lk}\big) - \ln \upsilon^*_{lk}\big]^2 + \Psi'\big(\upsilon^*_{lk}\big) \tag{3.67}$$

$$\overline{\sigma}_{l} = \frac{s^*_{l}}{t^*_{l}}, \left\langle \ln \sigma_{l} \right\rangle = \Psi\big(s^*_{l}\big) - \ln t^*_{l} \tag{3.68}$$

$$\left\langle \big(\ln \sigma_{l} - \ln \overline{\sigma}_{l}\big)^2 \right\rangle = \big[\Psi\big(s^*_{l}\big) - \ln s^*_{l}\big]^2 + \Psi'\big(s^*_{l}\big) \tag{3.69}$$

$$\overline{\tau}_{lk} = \frac{\Omega^*_{l}}{\Lambda^*_{l}}, \left\langle \ln \tau_{l} \right\rangle = \Psi\big(\Omega^*_{l}\big) - \ln \Lambda^*_{l} \tag{3.70}$$

$$\left\langle \big(\ln \tau_{l} - \ln \overline{\tau}_{l}\big)^2 \right\rangle = \big[\Psi\big(\Omega^*_{l}\big) - \ln \Omega^*_{l}\big]^2 + \Psi'\big(\Omega^*_{l}\big) \tag{3.71}$$

$$\left\langle z_{dnk} \right\rangle = \phi_{dnk}, \left\langle y_{dl} \right\rangle = r_{dl}, \left\langle \ln \beta_{kv} \right\rangle = \Psi(\lambda_{kv}) - \Psi(\sum_{f=1}^{V} \lambda_{kf}) \tag{3.72}$$

We follow the same algorithm for LBLMA calculating the equations 3.44 - 3.49 repeatedly until convergence.

3.3 Online Variational Inference

The variational algorithm introduced in the previous section of the chapter gives faster convergence than pure Bayesian approach with Gibbs sampling

in terms of batch data. However, we might encounter situations where the data is huge and as a result requires processing them in mini batches due to memory constraints or the data is constantly arriving in realtime. In these cases we use online variational inference algorithm which is more equipped to handle this type of data [67]. To modify the variational algorithm for online learning let us consider a part of the complete set of D documents. Let $p(W)$ be the model evidence of this finite set of documents. The expectation value of the model evidence is thus given by,

$$\big\langle \ln p(W) \big\rangle = \int \Phi(W) \ln \left(\int p(W \mid \Theta) p(\Theta) d(\Theta) \right) dW \tag{3.73}$$

where, the unknown probability distribution of the data observed until now is given by $\Phi(W)$. Then expectation of the lower bound considering the hyperparameter set $\Omega = \{\vec{\sigma}, \vec{\tau}, \vec{\beta}, \vec{\theta}\}$ in case of GD or $\Omega = \{\vec{\sigma}, \sigma, \tau, \vec{\beta}, \vec{\theta}\}$ in case of BL can then be modified as,

$$\begin{aligned}
\big\langle \mathcal{L}(Q) \big\rangle_{\Phi} =& \left\langle \sum_{Z} \sum_{y} \int Q(\Omega) Q(\mathcal{Z}) Q(\vec{y}) \ln \left[\frac{p(W, \mathcal{Z}, \vec{y} \mid \Omega) p(\Omega)}{Q(\Omega) Q(\mathcal{Z}) Q(\vec{y})} \right] d\Omega \right\rangle_{\Phi} \\
=& D \int Q(\Omega) d\Omega \left\langle \sum_{Z} \sum_{y} Q(\mathcal{Z}) Q(\vec{y}) \ln \left[\frac{p(W, \mathcal{Z}, \vec{y} \mid \Omega)}{Q(\mathcal{Z}) Q(\vec{y})} \right] \right\rangle_{\Phi} \\
&+ \int Q(\Omega) \ln \left[\frac{p(\Omega)}{Q(\Omega)} \right] d\Omega
\end{aligned} \tag{3.74}$$

If we take only a part of the data, let's say x documents among D, then the lower bound corresponding to this smaller set is given by,

$$\begin{aligned}
\mathcal{L}^{x}(Q) =& \frac{D}{x} \sum_{i=1}^{x} \int Q(\Omega) d\Omega \sum_{Z_i} \sum_{y_i} Q(\vec{Z}_i) Q(\vec{y}_i) \ln \left[\frac{p(W_i, \vec{Z}_i, \vec{y}_i \mid \Omega)}{Q(\vec{Z}_i) Q(\vec{y}_i)} \right] \\
&+ \int Q(\Omega) \ln \left[\frac{p(\Omega)}{Q(\Omega)} \right] d\Omega
\end{aligned} \tag{3.75}$$

We maximize the lower bound each time with the assumption that we have observed x documents, which is a subset of the entire corpus. As the new document $\vec{w_x}$ is streamed, we maximize the lower bound $\mathcal{L}^{x}(Q)$ with respect to $Q(\vec{z}_x)$ and $Q(\vec{y}_x)$ using the variational solution of the parameters in the

previous step $Q^{(x-1)}(\Omega)$ followed by maximizing the lower bound with respect to $Q^{(x)}(\Omega)$ by keeping $Q(\vec{y}_x)$ and $Q(\vec{Z}_x)$ fixed. This stochastic approach makes the algorithm a natural gradient method. The parameter updates for LGDMA from the batch variational approximation changes to,

$$\begin{aligned}
\Delta\upsilon_{lk}^{*(x)} =& \upsilon_{lk}^{*(x)} - \upsilon_{lk}^{*(x-1)} \\
\upsilon_{lk}^{*} =& \upsilon_{lk} + D\big\langle y_{xl}\big\rangle \Bigg[\Psi\big(\overline{\sigma}_{lk} + \overline{\tau}_{lk}\big) - \Psi\big(\overline{\sigma}_{lk}\big) \\
& + \overline{\tau}_{lk}\Psi'\big(\overline{\sigma}_{lk} + \overline{\tau}_{lk}\big)\big(\big\langle \ln \tau_{lk}\big\rangle - \ln \overline{\tau}_{lk}\big)\Bigg]\overline{\sigma}_{lk} - \upsilon_{lk}^{*(x-1)}
\end{aligned} \tag{3.76}$$

$$\begin{aligned}
\Delta s_{lk}^{*(x)} =& s_{lk}^{*(x)} - s_{lk}^{*(x-1)} \\
s_{lk}^{*} =& s_{lk} + D\big\langle y_{xl}\big\rangle \Bigg[\Psi\big(\overline{\tau}_{lk} + \overline{\sigma}_{lk}\big) - \Psi\big(\overline{\tau}_{lk}\big) \\
& + \overline{\sigma}_{lk}\Psi'\big(\overline{\tau}_{lk} + \overline{\sigma}_{lk}\big)\big(\big\langle \ln \sigma_{lk}\big\rangle - \ln \overline{\sigma}_{lk}\big)\Bigg]\overline{\tau}_{lk} - s_{lk}^{*(x-1)}
\end{aligned} \tag{3.77}$$

$$\begin{aligned}
\Delta\nu_{lk}^{*(x)} =& \nu_{lk}^{*(x)} - \nu_{lk}^{*(x-1)} \\
\nu_{lk}^{*} =& \nu_{lk} - D\big\langle y_{xl}\big\rangle\big\langle \ln \theta_{xk}\big\rangle - \nu_{lk}^{*(x-1)}
\end{aligned} \tag{3.78}$$

$$\begin{aligned}
\Delta t_{lk}^{*(x)} =& t_{lk}^{*(x)} - t_{lk}^{*(x-1)} \\
t_{lk}^{*} =& t_{lk} - D\big\langle y_{xl}\big\rangle \Bigg\langle \ln\Big[1 - \sum_{j=1}^{K}\theta_{xj}\Big]\Bigg\rangle - t_{lk}^{*(x-1)}
\end{aligned} \tag{3.79}$$

$$\begin{aligned}
\Delta g_{dk}^{*(x)} =& g_{dk}^{*(x)} - g_{dk}^{*(x-1)} \\
g_{dk}^{*} =& g_{dk} + D\sum_{n=1}^{N_d}\big\langle z_{dnk}\big\rangle + \sum_{l=1}^{L}\big\langle y_{dl}\big\rangle\sigma_{lk} - g_{dk}^{*(x-1)}
\end{aligned} \tag{3.80}$$

$$\begin{aligned}
\Delta h_{dk}^{*(x)} =& h_{dk}^{*(x)} - h_{dk}^{*(x-1)} \\
h_{dk}^{*} =& h_{dk} + \sum_{l=1}^{L}\big\langle y_{dl}\big\rangle\tau_{lk} + \sum_{kk=k+1}^{K}\phi_{dn(kk)} - h_{dk}^{*(x-1)}
\end{aligned} \tag{3.81}$$

$$\Delta\lambda_{kv}^{*(x)} = \lambda_{kv}^{*(x)} - \lambda_{kv}^{*(x-1)}$$
$$\lambda_{kv}^{*} = \lambda_{kv} + D\sum_{n=1}^{N_d}\sum_{v=1}^{V}\phi_{xnk}w_{xnv} - \lambda_{kv}^{*(x-1)} \tag{3.82}$$

$$\Delta = \pi_{l}^{*(x)} - \pi_{l}^{*(x-1)} = \frac{D}{x}r_{tl} - \pi_{l}^{*(x-1)} \tag{3.83}$$

$$\upsilon_{lk}^{*(x)} = \upsilon_{lk}^{*(x-1)} + \varpi_x\Delta\upsilon_{lk}^{*(x)} \qquad \nu_{lk}^{*(x)} = \nu_{lk}^{*(x-1)} + \varpi_x\Delta\nu_{lk}^{*(x)} \tag{3.84}$$
$$s_{lk}^{*(x)} = s_{lk}^{*(x-1)} + \varpi_x\Delta s_{lk}^{*(x)} \qquad t_{lk}^{*(x)} = t_{lk}^{*(x-1)} + \varpi_x\Delta t_{lk}^{*(x)} \tag{3.85}$$
$$g_{dk}^{*(x)} = g_{dk}^{*(x-1)} + \varpi_x\Delta g_{dk}^{*(x)} \qquad h_{dk}^{*(x)} = h_{dk}^{*(x-1)} + \varpi_x\Delta h_{dk}^{*(x)} \tag{3.86}$$
$$\lambda_{kv}^{*(x)} = \lambda_{kv}^{*(x-1)} + \varpi_x\Delta\lambda_{kv}^{*(x)} \qquad \pi_{l}^{*(x)} = \pi_{l}^{*(x-1)} + \varpi_x\Delta\pi_{l}^{*(x)} \tag{3.87}$$

with the learning rate ϖ_x given by $\varpi_x = (\mu_0 + x)^{-\varepsilon}$ following the constraints $\varepsilon \in (0.5, 1]$ and $\mu_0 \geq 0$ as mentioned in [9]. This helps reduce the effects of estimation from past documents over time. These equations are calculated every time new data arrives.

Similarly, we can also write the corresponding updates for LBLMA model as,

$$\Delta\upsilon_{lk}^{*(x)} = \upsilon_{lk}^{*(x)} - \upsilon_{lk}^{*(x-1)}$$
$$\upsilon_{lk}^{*} = \upsilon_{lk} + D\sum_{d=1}^{D}\langle y_{dl}\rangle\overline{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K}\overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk}) + \Psi\Big(\sum_{k=1}^{K}\Big)\sum_{a\neq k}^{K}\big(\langle\ln\mu_{la}\rangle - \ln\overline{\mu}_{la}\big)\overline{\mu}_{la}\Big] - \upsilon_{lk}^{*(x-1)} \tag{3.88}$$

$$\Delta\nu_{lk}^{*(x)} = \nu_{lk}^{*(x)} - \nu_{lk}^{*(x-1)}$$
$$\nu_{lk}^{*} = \nu_{lk} - D\sum_{d=1}^{D}\langle y_{dl}\rangle\Big[\langle\ln\theta_{dk}\rangle - \Big\langle\ln\sum_{k=1}^{K}\theta_{dk}\Big\rangle\Big] - \nu_{lk}^{*(x-1)} \tag{3.89}$$

$$
\begin{aligned}
\Delta s_l^{*(x)} =& s_l^{*(x)} - s_l^{*(x-1)} \\
s_l^* =& s_l + D\sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Bigg[\Psi\big(\overline{\sigma}_l+\overline{\tau}_l\big)-\Psi\big(\overline{\sigma}_l\big) \\
&+\overline{\tau}_l\Psi'\big(\overline{\sigma}_l+\overline{\tau}_l\big)\big(\big\langle\ln\tau_l\big\rangle-\ln\overline{\tau}_l\big)\Bigg]\overline{\sigma}_l - s_l^{*(x-1)}
\end{aligned}
\tag{3.90}
$$

$$
\begin{aligned}
\Delta t_l^{*(x)} =& t_l^{*(x)} - t_l^{*(x-1)} \\
t_l^* =& t_l - D\sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Bigg\langle\ln\Big[\sum_{k=1}^{K}\theta_{dk}\Big]\Bigg\rangle - s_l^{*(x-1)}
\end{aligned}
\tag{3.91}
$$

$$
\begin{aligned}
\Delta\Omega_l^{*(x)} =& \Omega_l^{*(x)} - \Omega_l^{*(x-1)} \\
\Omega_l^* =& \Omega_{lk} + D\sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Bigg[\Psi\big(\overline{\tau}_l+\overline{\sigma}_l\big)-\Psi\big(\overline{\tau}_l\big) \\
&+\overline{\sigma}_l\Psi'\big(\overline{\tau}_l+\overline{\sigma}_l\big)\big(\big\langle\ln\sigma_l\big\rangle-\ln\overline{\sigma}_l\big)\Bigg]\overline{\tau}_l - \Omega_l^{*(x-1)}
\end{aligned}
\tag{3.92}
$$

$$
\begin{aligned}
\Delta\Omega_l^{*(x)} =& \Omega_l^{*(x)} - \Omega_l^{*(x-1)} \\
\Lambda_l^* =& \Lambda_l - D\sum_{d=1}^{D}\big\langle y_{dl}\big\rangle\Bigg\langle\ln\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]\Bigg\rangle - \Omega_l^{*(x-1)}
\end{aligned}
\tag{3.93}
$$

$$
\begin{aligned}
\Delta f_{dk}^{*(x)} =& f_{dk}^{*(x)} - f_{dk}^{*(x-1)} \\
f_{dk}^* =& f_{dk} + D\sum_{n=1}^{N_d}\big\langle z_{dnk}\big\rangle + D\sum_{l=1}^{L}\big\langle y_{dl}\big\rangle\sigma_{lk} - f_{dk}^{*(x-1)}
\end{aligned}
\tag{3.94}
$$

$$
\begin{aligned}
\Delta g_d^{*(x)} =& g_d^{*(x)} - g_d^{*(x-1)} \\
g_d^* =& g_d + D\sum_{n=1}^{N_d}\sum_{k=1}^{K}\big\langle z_{dnk}\big\rangle + D\sum_{l=1}^{L}\big\langle y_{dl}\big\rangle\sigma_l - g_d^{*(x-1)}
\end{aligned}
\tag{3.95}
$$

$$\begin{aligned}\Delta h_d^{*(x)} =& h_d^{*(x)} - h_d^{*(x-1)} \\ h_d^* =& h_d + D\sum_{l=1}^{L}\langle y_{dl}\rangle\tau_l - h_d^{*(x-1)}\end{aligned} \tag{3.96}$$

$$\begin{aligned}\Delta\lambda_{kv}^{*(x)} =& \lambda_{kv}^{*(x)} - \lambda_{kv}^{*(x-1)} \\ \lambda_{kv}^* =& \lambda_{kv} + D\sum_{n=1}^{N_d}\sum_{v=1}^{V}\phi_{xnk}w_{xnv} - \lambda_{kv}^{*(x-1)}\end{aligned} \tag{3.97}$$

$$\Delta = \pi_l^{*(x)} - \pi_l^{*(x-1)} = \frac{D}{x}r_{tl} - \pi_l^{*(x-1)} \tag{3.98}$$

$$\upsilon_{lk}^{*(x)} = \upsilon_{lk}^{*(x-1)} + \varpi_x\Delta\upsilon_{lk}^{*(x)} \qquad \nu_{lk}^{*(x)} = \nu_{lk}^{*(x-1)} + \varpi_x\Delta\nu_{lk}^{*(x)} \tag{3.99}$$

$$s_l^{*(x)} = s_l^{*(x-1)} + \varpi_x\Delta s_l^{*(x)} \qquad t_l^{*(x)} = t_l^{*(x-1)} + \varpi_x\Delta t_l^{*(x)} \tag{3.100}$$

$$\Omega_l^{*(x)} = \Omega_l^{*(x-1)} + \varpi_x\Delta\Omega_l^{*(x)} \qquad \Lambda_l^{*(x)} = \Lambda_l^{*(x-1)} + \varpi_x\Delta\Lambda_l^{*(x)} \tag{3.101}$$

$$f_{dk}^{*(x)} = f_{dk}^{*(x-1)} + \varpi_x\Delta f_{dk}^{*(x)} \qquad g_d^{*(x)} = g_d^{*(x-1)} + \varpi_x\Delta g_d^{*(x)} \tag{3.102}$$

$$h_d^{*(x)} = h_d^{*(x-1)} + \varpi_x\Delta h_d^{*(x)} \qquad \lambda_{kv}^{*(x)} = \lambda_{kv}^{*(x-1)} + \varpi_x\Delta\lambda_{kv}^{*(x)} \tag{3.103}$$

$$\pi_l^{*(x)} = \pi_l^{*(x-1)} + \varpi_x\Delta\pi_l^{*(x)} \tag{3.104}$$

With these updates, we can achieve an efficient online learning for LBLMA.

3.4 Supervised models

The two algorithms mentioned before help in extracting topics from the data. However, the topic learned is better evaluated when used in combination with a classifier for a real world task. There are a lot of studies to use the learned topics from LDA as a feature space which can be used as input to a classifier like SVM, naive Bayes, etc [68]. In our paper, we use simultaneous learning of topics corresponding to the classes as mentioned in [10]. Figs. 3.3 and 3.4 show the graphical representation of our supervised LGDMA (sLGDMA) and supervised LBLMA (SLBLMA) model respectively. Let C be the number of classes in the dataset indicated by the association vector $\vec{c} = \{c_d\}$. We use a softmax function to define the class corresponding to a document $\vec{w}_d$

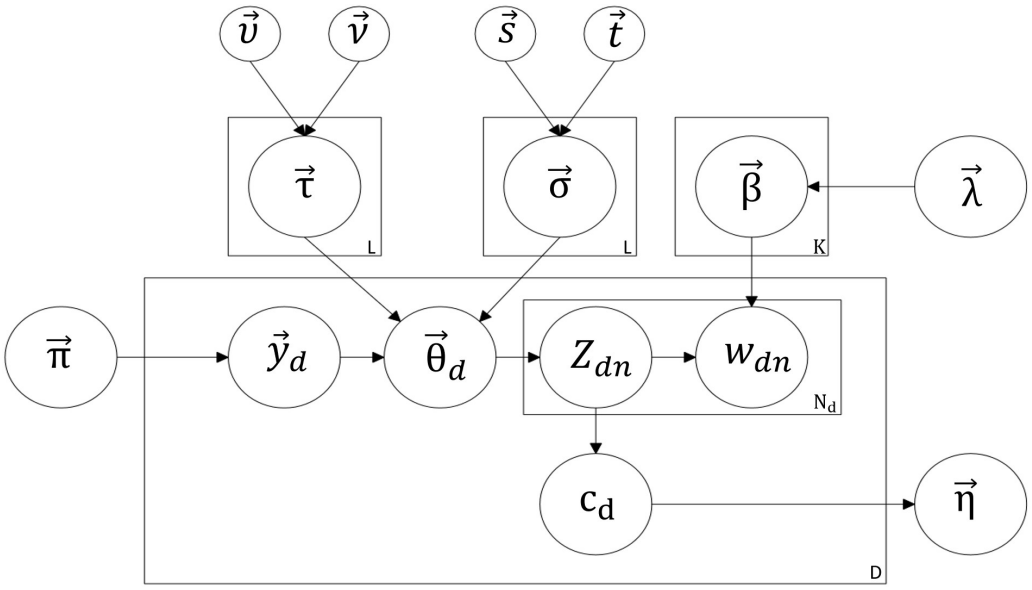

Figure 3.3: Graphical representation of supervised LGDMA

as mentioned in [69] given by,

$$p(c_d \mid \vec{z}_d, \vec{\eta}) = softmax(\eta_a^T \vec{z}_d) = \frac{\exp(\eta_a^T \vec{z}_d)}{\sum_{a=1}^{C} \exp(\eta_a^T \vec{z}_d)} \tag{3.105}$$

where $\vec{z}_d = \{\overline{z}_{dk}\}$ and $\vec{\eta} = \{\vec{\eta}_a\} = \{\eta_{ak}\}$ represents the class label coefficients which is more like a weight vector for each of the topics pertaining to the class. $\overline{z}_{dk} = \frac{1}{N_d} \sum_{n=1}^{N_d} \delta(z_{dn}, k)$ with $\delta(z_{dn}, k) = 1$ when $z_{dn} = k$ and 0 if not. The lower bound of the batch variational algorithms for both the models remains the same except for the addition of the softmax function. The softmax function impacts only the variable ϕ_{dnk} and hence the rest of the variational solutions remain untouched. Furthermore, there is also an addition of variational solution corresponding to the parameter $\vec{\eta}$. The new estimation of δ_{dnk} which causes the changes in ϕ_{dnk} is given by,

$$\hat{\delta}_{dnk} = \exp\Big(\big\langle \ln \beta_{kv} \big\rangle + \big\langle \ln \theta_{dk} \big\rangle + \frac{\eta_{ak}}{N_d} - (\vec{h}^T \vec{\phi}_{dn}^{old})^{-1} h_k\Big) \tag{3.106}$$

Given, $\vec{h} = \{h_k\}$ and

$$\vec{h}^T \vec{\phi}_{dn}^{old} = \sum_{a=1}^{C} \prod_{n=1}^{N_d} \Big(\sum_{j=1}^{K} \phi_{dnj} \exp\big(\frac{\eta_{aj}}{N_d}\big)\Big) \tag{3.107}$$

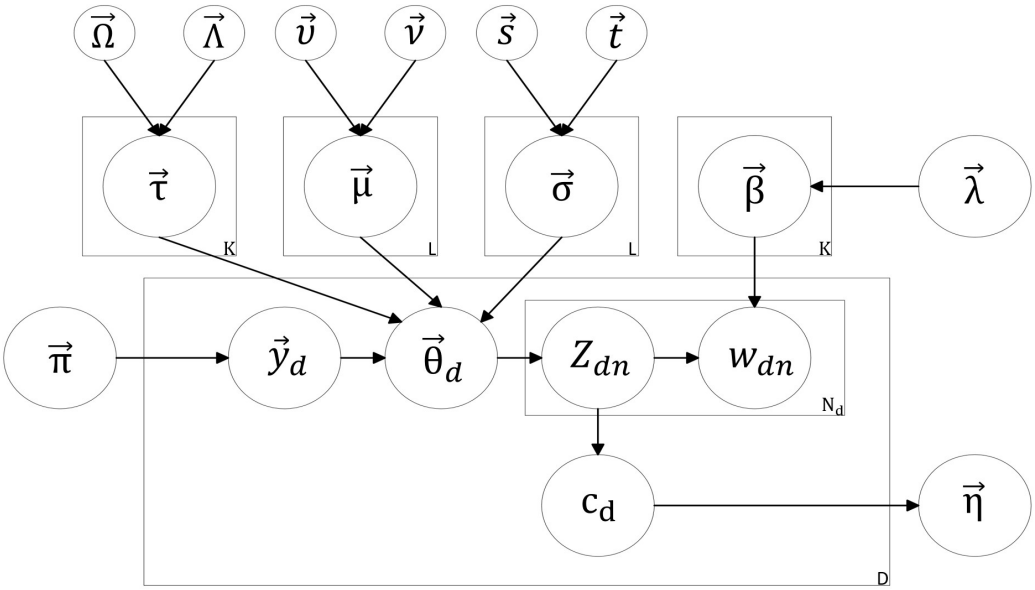

Figure 3.4: Graphical representation of supervised LBLMA

where $\vec{\phi}_{dn}^{old}$ is the value of ϕ_{dnk} from the previous iteration. We also use the estimation of $\vec{\eta}$ as in [69] given by,

$$\frac{\partial \mathcal{L}(\vec{\eta})}{\partial \eta_{ak}} = \sum_{d=1}^{D} \left[\delta(c_d, a)\vec{\phi}_{dk} - \left[\kappa_d^{-1} \prod_{n=1}^{N_d} \Big(\sum_{j=1}^{K} \phi_{dnj} \exp\big(\frac{\eta_{aj}}{N_d}\big) \Big) \right] \times \frac{\frac{1}{N_d}\phi_{dnk} \exp(\frac{\eta_{ak}}{N_d})}{\sum_{j=1}^{K} \phi_{dnk} \exp(\frac{\eta_{ak}}{N_d})} \right] \tag{3.108}$$

with,

$$\kappa_d = \sum_{a=1}^{C} \prod_{n=1}^{N_d} \Big(\sum_{k=1}^{K} \phi_{dnj} \exp\big(\frac{\eta_{ak}}{N_d}\big) \Big) \tag{3.109}$$

$$\vec{\phi}_{dk} = \frac{1}{N_d} \sum_{n=1}^{N_d} \phi_{dnk} \tag{3.110}$$

Detailed derivations of these equations are explained in [69]. During the training phase, either the batch variational algorithm or the online version (depends on the application under consideration) can be used with the new

solution for ϕ_{dnk} and $\vec{\eta}$ is calculated as an extra parameter in each iteration. For predicting a new document w_d, the variational estimations of a fitted model can be used to calculate $\vec{z}_d$ with the normal variational algorithm. We then calculate the value of $softmax(\eta_a^T \vec{z}_d)$ for each class and the class with the maximum value is chosen as the target label for the test document.

3.5 Experimental Results

The best way to evaluate topic models is to test its efficiency for classification tasks. Though the use of topic models is mostly in the field of natural language processing it has been widely deployed for multiple applications in other fields. Image categorization for example is an interesting application that has been surveyed extensively as well [69]. Genomic sequence classification is a field which is less explored with LDA. In our experiments, we evaluate our model against three applications. To prove the robustness of our model to extract topics in cases where the number of data samples are less, we use minimal data for training in our applications. We compare our model with other standard benchmark classifiers which uses topics learned from an LDA model. We also analyze the performance of our model when the number of topics and number of mixture components are changed. Section 3.5.1 shows how LGDMA model can be used to classify Genomic sequences of bacteria belonging to different genera. In section 3.5.2, we see how the model performs in categorizing images. Section 3.5.3 presents the results on text classification both for batch processing and our online model.

Table 3.1: Accuracy of classification models on different applications

Model	Genome(%)	Image(%)	Text(%)
LDA + SVC	84.28	82.5	94.2
LDA + KNN	82.85	75	93.8
LDA + RF	83.57	82.5	93.8
sLGDMA	86.43	80.83	95.6
sLBLMA	87.86	83.33	96.4

3.5.1 Genomic Sequence Classification

Classifying a new genomic sequence which might be DNA or RNA into the proper species requires comparing the sequence with around 9 million sequences. Classifying them into their taxonomic domain, phylum, class, order, etc helps narrowing down the search area which speeds up the process. Probabilistic topic models like LGDMA and LBLMA can be of help for this task [70]. Any living organism can be identified by its DNA or RNA. RNA consists of four nucleotides namely, adenine (A), guanine (G), cytosine (C) and uracil (U). DNA on the other hand varies in the last component which is thymine (T). The sequence of these compositions vary for each species and hence can be identified uniquely.

For our experiments, we use the dataset provided in [71]. The data consists of genomic sequences from 3 domains 'Bacteria', 'Archaea' and 'Eukaryota'. Since classifying between domains or even families is going to be a pretty easy task, we choose 7 different genera from bacteria namely 'Kocuria', 'Arthrobacter', 'Micrococcus', 'Pseudarthrobacter', 'Rothia' 'Glutamicibacter' and 'Paenarthrobacter' from the family 'Micrococcaceae' consisting of 120 genomic sequences each. Among them 100 is used for training and 20 for testing. In the preprocessing, considering a sequence of any length, we extract k-mers and count the frequency of each k-mer in the sequence. This gives a similar representation to the bag of words model. A k-mer is a string of length k in a genomic sequence. For example, if $k = 7$ we extract all possible series of length 7 from the sequence. We use this count data as input for our supervised LGDMA and LBLMA model.

Fig. 3.5 and 3.5 shows how changing the number of topics K and the number of mixture components L affects the performance of our models. In the case supervised LGDMA, we can see that maximum accuracy is achieved when $K = 25$ whereas supervised LBLMA performs the best with lower number of topics of just $K = 15$. Both models perform the best when the number of components is set to $L = 3$. We also see that the sLBLMA performs better than the sLGDMA model. However, generally both these models perform much better than the other standard classifiers based on support vector machine (SVM), K-nearest neigbors (KNN) and random forest (RF) as shown in Fig. 3.1. The input for these classifiers will be the output parameter of the topic proportions from LDA as described in [2]. It is a notable fact that when $L = 1$ the model is technically LGDA and LBLA models with no impact from the mixture components. This comparison shows

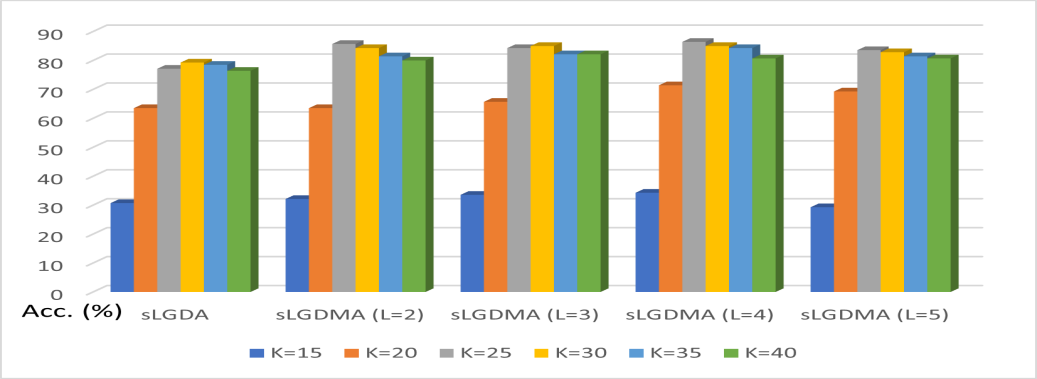

Figure 3.5: Variations of accuracy over L and K for genome classification supervised LGDMA

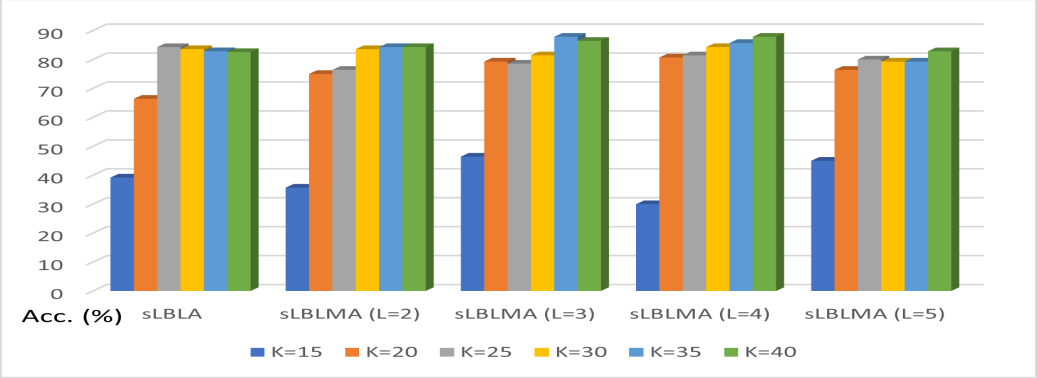

Figure 3.6: Variations of accuracy over L and K for genome classification supervised LBLMA

the benefits of adding mixture components to the topic model.

3.5.2 Image Classification

Image classification is an important application in the field of pattern recognition. Since image data can be represented as a bag of visual words, it is a known fact that topic models such as sLGDMA and sLBLMA can be used for learning from this data. There are a number of methods which we can use to extract features from images before converting them into bags of visual words. For example, histogram of oriented gradients (HOG) is a method

which splits the image into cells and creates a histogram out of the gradient directions within the cell [72]. Scale invariant feature transform (SIFT) is another method which identifies a set of interest points and records the features corresponding to that point [73]. We then used these features to create a bag of visual words model which can be used as input for image classification.

Figure 3.7: Sample images from each of the class in GHIM dataset

In our experiments, we use the data from GHIM-10K dataset mentioned in [74]. The dataset has 500 images in each of its 20 categories with size 400×300 or vice versa. For our experiments, we use 6 classes from the dataset namely buildings, cars, flowers, planes, sail boats and chicken. Some sample images from the dataset are shown in figure 3.7. We take 100 images from each of this classes for training and 20 for testing. We extract SIFT feature descriptors from these images and then use k-means algorithm to create a bag of visual words model counting the frequency of similar features in the dataset. This data is served as input for our sLGDMA and sLBLMA models. Fig. 3.8 and 3.9 shows the result obtained for a test set of 20 images each. We can see from table 3.1 that sLBLMA achieves the best performance. However, sLGDMA performs comparatively less efficient compared to other models. This experiment stands as an example to show why we need sLBLMA to overcome the drawbacks of LGDMA. The best results for sLBLMA in this experiment was achieved at $K = 15$ and $L = 3$. For sLGDMA the best results where when K is set to 25 and L is 3. In our experiments, we also found the pattern that as L increases, the accuracy increases up to a certain point which is 3 in our case and then decreases. This is a similar effect usually observed with the number of topics K.

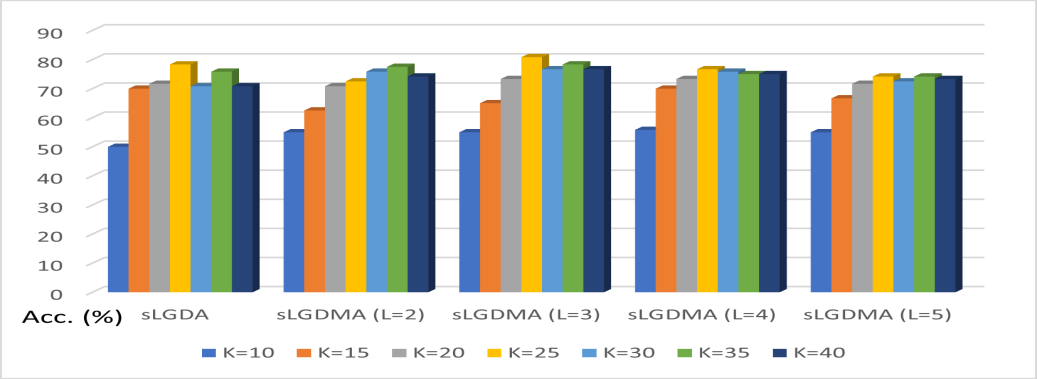

Figure 3.8: Variations of accuracy over L and K for Image classification with supervised LGDMA

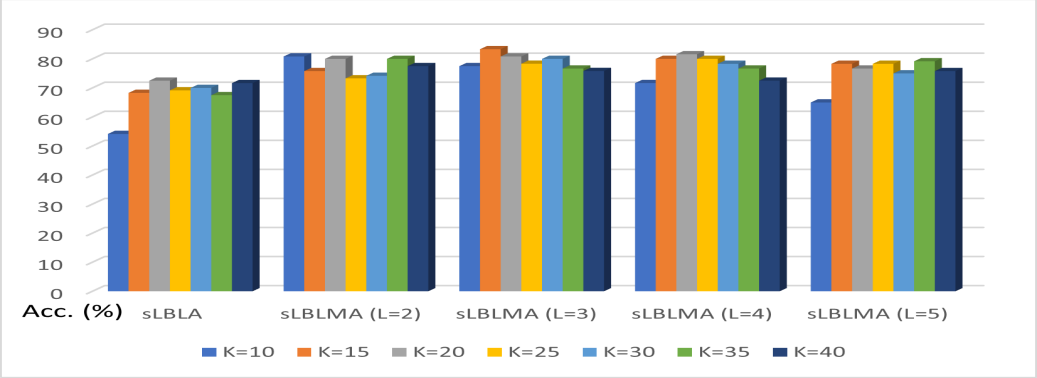

Figure 3.9: Variations of accuracy over L and K for Image classification with supervised LBLMA

3.5.3 Text Classification

Text classification is primarily the basic application that topic models like LGDMA were invented for and allows us to visualize the topics which are understandable. For text classification, we use BBC news data[1] belonging to five categories which are business, entertainment, politics, sport and technology . The dataset consists of about 2225 documents with 610 documents from business, 386 documents concerning entertainment, 417 documents from politics, 511 documents from sports and 401 documents from technology re-

[1]http://mlg.ucd.ie/datasets/bbc.html

lated documents. Among these documents 100 from each of the classes was separated for testing and the rest were used for training. The preprocessing step involved removal of stop words followed by removing small words less than the length of 4. Both lemmatization and stemming steps was ignored as they did not give good results. The frequency of the words in the tokenized documents is recorded to form the bag of words representation. The data is then fed into our sLGDMA and sLBLMA models for classification. We show the topics learned by our model corresponding to LGDMA in Table 3.2 as an example. Similar topics where recorded with LBLMA as well which is avoided for brevity. This shows the interpretability of the model and all the learned topics seems fairly relatable. Table 3.1 shows that both our models

Table 3.2: Accuracy of models on text data

Topic	Words
Business	'sales','growth','firm','economic','economy','government', 'market','company'
Entertainment	'singer','actor','album','star','band','award', 'awards','music'
Politics	'leader','secretary','prime','said','minister','brown', 'party','blair'
Sports	'players','team','second','play','world','game', 'time','england'
Technology	'video','game','phone','digital','software','games', 'users','music'

perform really well compared to other standard models. We can also see how the variations in number of topics affects the results in Fig. 3.10 and 3.11. sLGDMA seems to perform the best when $K = 30$ and $L = 3$. On the other hand, the accuracy of sLBLMA is not impacted too much for $K = 30$ for L from 2 to 4 though it is not the case for different values of K.

Being the largest dataset in our experiments, this application serves as an excellent candidate to test our online models. We run the online versions of LGDMA and LBLMA on this data with settings from the best performances in batch experiments. The results are shown in Fig. 3.12 which presents how the accuracy varies when the algorithm has seen every 100 documents. We can see that sLGDMA takes more than 200 documents to properly fit

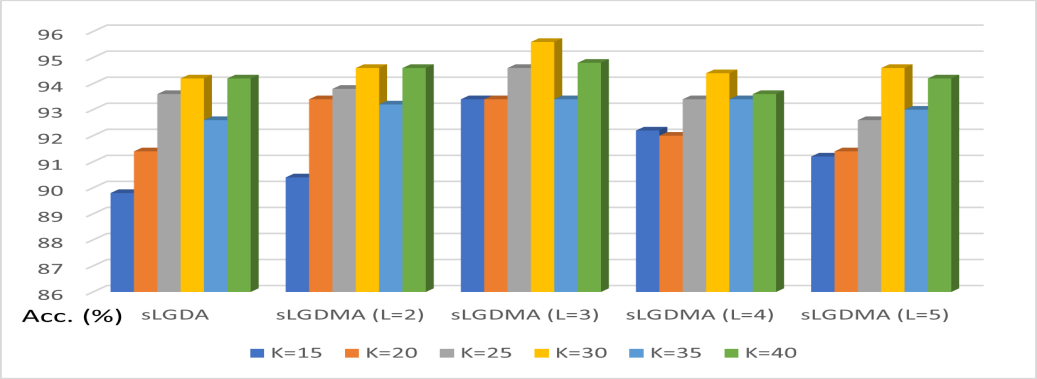

Figure 3.10: Variations of accuracy over L and K for text classification supervised LGDMA

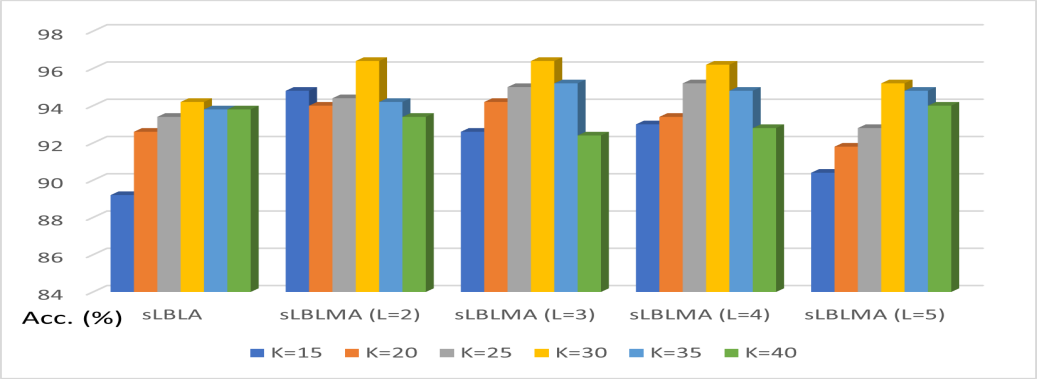

Figure 3.11: Variations of accuracy over L and K for text classification supervised LBLMA

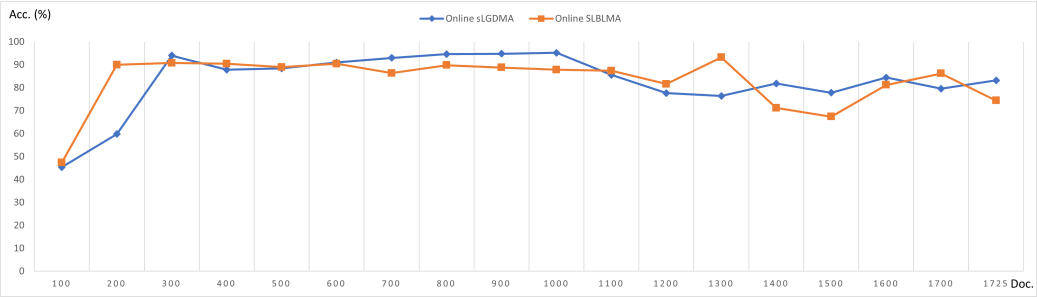

Figure 3.12: Performance of online sLGDMA and sLBLMA

the document whereas sLBLMA was able to get around 90% accuracy after learning 200 documents. We do not achieve high results as in the batch variational approach because the order in which the documents are learned plays a pivotal role in the accuracy. The best accuracy achieved by sLGDMA was 95.2% after seeing 1000 documents and 93.2% in the case of sLBLMA. The average accuracy of sLGDMA is 82.8% and that of sLBLMA is 82.97%. Though slightly higher we can see that sLGDA tends to be me more stable over time. However, it is obvious that both models present their own set of trade offs which dictates the choice of what model is to be used for a particular application.

Chapter 4

Interactive Learning

A lot of efforts have been put in recent times for research in the field of natural language processing. One of the major tasks in natural language processing is to categorize texts into different categories. Extracting topics is undoubtedly one of the most important tasks in this area of research. This helps in various tasks such as sentiment analysis [75], threat detection [76], document categorization [77], etc. In the previous chapter we proposed the use of LGDMA and LBLMA models for topic extraction which could help in text categorization. However, there is a chance that the topics derived may contain words which are irrelevant to that topic.

Improving the quality of topics extracted from these models is important for accurate inference and unsupervised language tasks. Owing to this cause, in this chapter, we propose interactive latent generalized Dirichlet Liouville mixture allocation (iLGDMA) and interactive Beta-Liouville mixture allocation (iLBLMA) models which combines the clustering capabilities with interactive learning which helps the user to modify the topic weights of irrelevant words within the topic. The experiments were done separately for iLGDMA and iLBLMA models on different datasets since the efficiency of LGDMA and LBLMA have been already proven in the previous chapter.

Section 4.1 introduces our approach of interactive learning and explains how the basic mixture allocation models can be easily adapted for interactive learning. The experiments performed and the results obtained are discussed separately for iLGDMA and iLBLMA in sections 4.2 and 4.3 respectively.

4.1 Mixture Allocation Models with Interactive Learning

Recollecting from the previous chapter, with variational smoothing the topic word proportions is given by Eq. 3.8. For the sake of interactive learning, the topic words parameter $\vec{\beta}$ is split into the objective variable $\vec{\beta}_o$ which represents the probability generated by the iLGDMA or iLBLMA model and $\vec{\beta}_u$ is the user defined subjective probabilities decided by a human expert. The primal motive to incorporate an interactive learning algorithm is to enhance topic quality, by asking the users to decide the probability of words within the topic. Using the definitions of $\vec{\beta}_o$ and $\vec{\beta}_u$ mentioned earlier and adding weights to this objective and subjective probabilities, $\vec{\beta}$ can be defined as.

$$\vec{\beta} = \eta_1 \vec{\beta}_o + \eta_2 \vec{\beta}_u \tag{4.1}$$

where, η_1 and η_2 are the weights given to the objective and subjective probabilities respectively. The intention of using these values is to control the effect of user input as it is probable that in some cases the user might not be well versed in the topic under consideration. So assigning a lower value to η_2 in this case helps us to reduce the impact of user defined input. For example, If the user has sufficient knowledge in the subject then η_1 can be for example 0.2 and η_2 can be 0.8. If the user doesn't have enough knowledge in the subject the values can be vice versa. The only criteria to be taken care of here is that $\eta_1 + \eta_2 = 1$. Apart from this change in the definition of $\vec{\beta}$, all the definitions and solutions from the previous chapter holds and the same variational procedure can be followed.

Another important question is to identify when to prompt users for input. The criteria can be custom-defined based on our needs and the problem we are trying to solve. For example, the criteria could be the number of iterations or when the coherence score beyond a certain value or when the classification accuracy with the topics is above a certain level. Nevertheless, the topics so derived by the model might have some inconsistencies lowering the quality of the topics. The main advantage of our models is going to be the interactive learning part which helps us improve the quality of topics so formed. Let's say that our model gives us K topics at convergence. We calculate the UMass coherence score at this point as explained in section 2.7.2.

Our new algorithm hence involves running the default variational inference to obtain the topics from the data and then assigning lower probabilities

to words which the user feels do not belong to that topic. At our convergence criteria, we prompt the users to modify the probabilities of T_k words within each topic. It would be better to show the topics with low coherence score based on a threshold in case of large number of topics. If $\{p^k_{o1}, p^k_{o2}, ..., p^k_{oT}\}$ are the probabilities inferred by our model and $\{p^k_{u1}, p^k_{u2}, ..., p^k_{uT'}\}$ is the set of probabilities of T'_k words that the user choose to modify, then the new probabilities for this set of words is obtained by $\beta^K_{ut} = p^k_{ot} * p^k_{ut}$. The probabilities that had been reduced is distributed among the rest of the words proportionally by using the equation:

$$\beta^k_{ut} = p_{ot} * \big(1 + \frac{p_r}{\sum_t p_{ot}}\big); \forall t \notin T'_k \tag{4.2}$$

Here, $p_r = \sum_t^{T'_k} p^k_{ot} * \big(1 - p^k_{ut}\big)$. The value of $\vec{\beta}_u$ can be substituted in Eq. 4.1. The variational algorithm is again carried out to attain the new set of words in each topic. The graphical representation of the models after the modification is shown in Fig. 4.1 and 4.2.

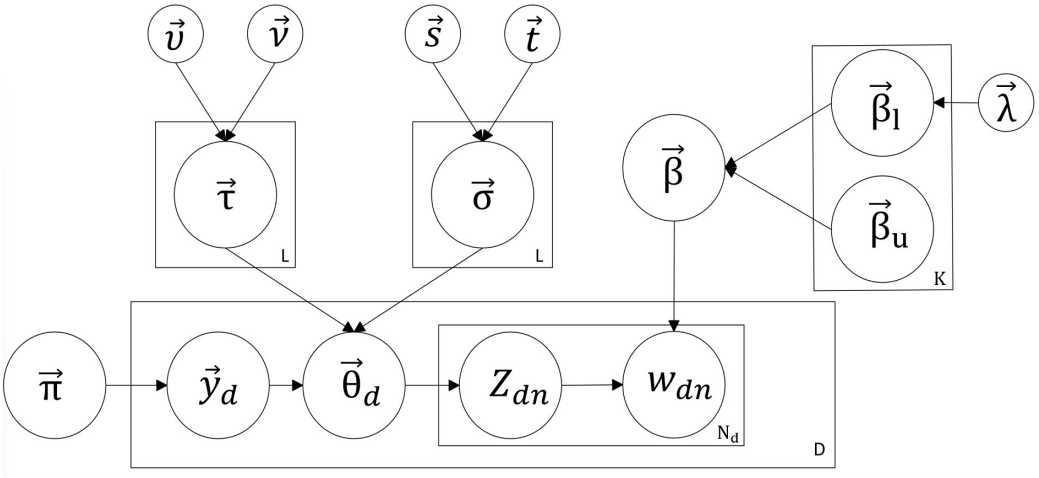

Figure 4.1: Graphical representation of iLGDMA

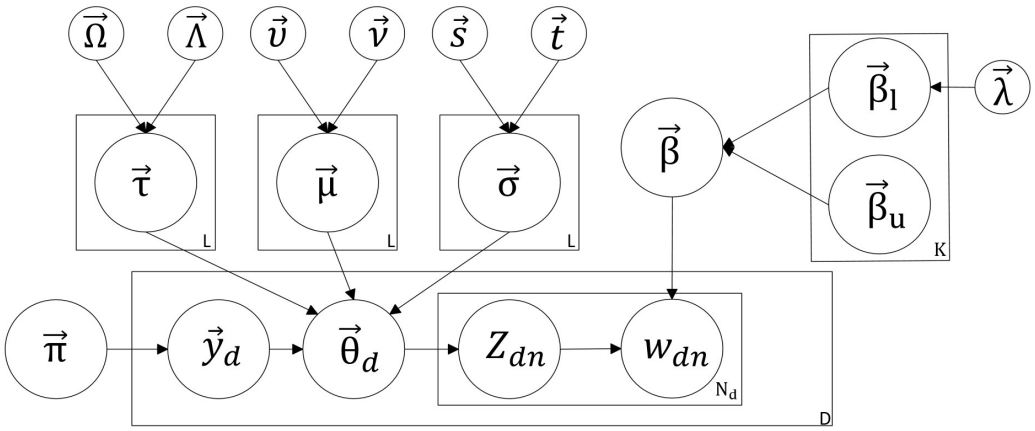

Figure 4.2: Graphical representation of iLBLMA

4.2 Experimental Results for iLGDMA

To see how our model performs with real data, we use two real world datasets namely, BBC news[1] and twitter emotions [78]. The former consists of 2325 documents from 5 categories specifically business (610), entertainment (386), politics (417), sports (511) and technology (401). On the other hand, the later entails a huge corpus 416,809 tweets representing emotions related to anger, joy, fear, love, sadness and surprise. To keep things simple we choose 2000 samples form each emotion category to test our model. The criteria in our case will be, to pause when the accuracy of a supervised version of LGDMA model following [10] attains a threshold accuracy which can be varied. At this point the user will be prompted to enter new probabilities for the words in each topic. After removing the stop words and words less than four letters we create a bag of words model with 1800 and 1000 most frequent words as the vocabulary. UMass score explained in section 2.7.2 is used as a metric to record the performance of our model.

We compare our model with vanilla LGDMA, latent generalized Dirichlet allocation (LGDA) and LDA models respectively. Our main comparison is between iLGDMA and LGDMA since the interactive version is a direct improvement over LGDMA. For both the experiments the value of L was found to give better results when set to 3. The rest of the parameters are randomly initiated. The coherence score of the two models when the value

[1]http://mlg.ucd.ie/datasets/bbc.html

of K is set as 10, 15, 20 and 25 is shown in table 4.1 and table 4.2 for the two datasets respectively. For both the cases, we see that iLGDMA achieves a better coherence score than rest of the bunch.

Table 4.1: Average coherence score of all topics for BBC news data

Model	K=10	K=15	K=20	K=25
iLGDMA	-1.04	-1.16	-1.20	-1.56
LGDMA	-1.20	-1.18	-1.36	-1.73
LGDA	-1.30	-1.27	-1.38	-1.77
LDA	-1.22	-1.40	-2.08	-1.87

Table 4.2: Average coherence score of all topics for emotions data

Model	K=10	K=15	K=20	K=25
iLGDMA	-4.65	-4.87	-4.67	-5.52
LGDMA	-4.86	-4.99	-5.90	-6.30
LGDA	-4.99	-5.09	-5.94	-7.48
LDA	-5.13	-6.17	-6.30	-7.67

However, it would be more appropriate to compare what percentage of increase in topic coherence our model provides compared to the rest, especially LGDMA. Figure 4.3 shows the percentage increase in coherence score using our iLGDMA with the other models for BBC news data. The figure also provides a fine comparison between the percentage increase achieved by varying the weights of user defined probabilities. The experiments were conducted with objective and subjective properties set to 0.2 and 0.8 to simulate a well versed user and vice-versa in case of a user with low subject knowledge. We can see a clear improvement when the user is someone who knows the subject as opposed to a mundane user. This was replicated by lowering the probability of a few correctly identified words in a topic in case of a common user. The figure shows that even in this case we can still see considerable improvements over the other models. This shows the robustness of our model. We can observe similar results in the case of tweets labelled with emotions. Another notable observation from the two experiments is that, though the weights were varied when the value of K was set to 10, the coherence score remained the same. This is because, most of the words in the topics were

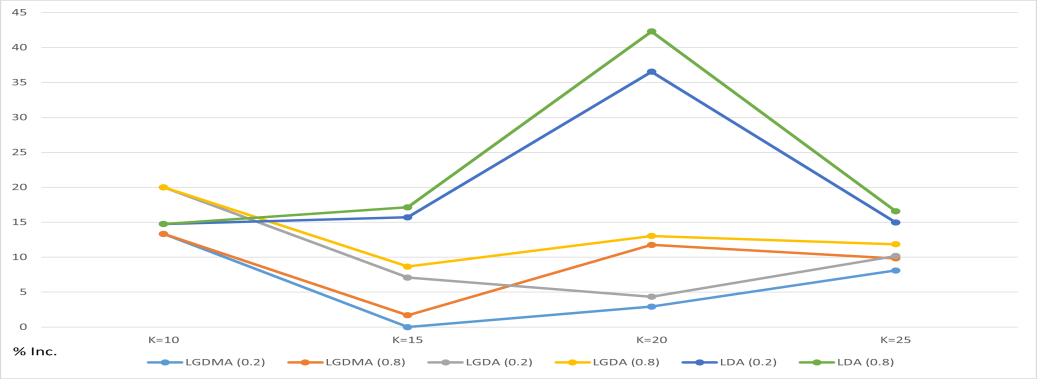

Figure 4.3: % Increase in topic quality for BBC news

already closely related and little information from the user was more than enough to improve the coherence to the best possible value. The figures also show that our model outputs better topics almost with a percentage increase of at least about 25% with respect to LDA for certain K values.

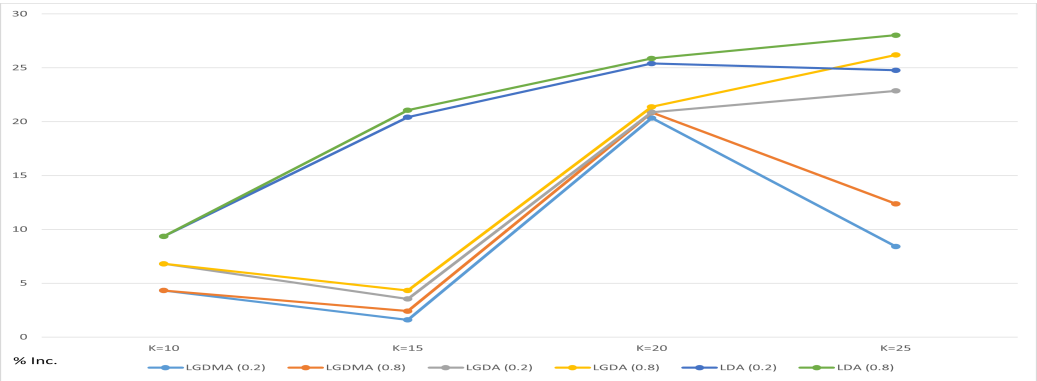

Figure 4.4: % Increase in topic quality for emotions dataset

4.3 Experimental Results for iLBLMA

We validate our model against the same BBC news data and 20 newsgroups[2]. 20 newsgroups data consists of about 20 different categories from various domains such as computers, recreation, science, sales, politics and religion. The preprocessing for both datasets involved removing the stop words and

[2]http://qwone.com/ jason/20Newsgroups/

words less than four letters in length. For BBC news we used a vocabulary of 1800 words and for 20 newsgroups, a vocabulary of about 5000 words were used. The initialization for the model was done randomly. In our experiments the best accuracy by using supervised iLGDMA is 96.4% and 75.69% for BBC and 20 newsgroups. Since topic quality is the prime concern with our model we compare the UMass coherence score with vanilla latent Beta-Liouville mixture allocation (LBLMA), latent Beta-Liouville allocation (LBLA) and LDA. The experiments were performed by varying the value of K from 10 to 25 in steps of 5. The best performance was achieved when the number of mixtures components L in our model was set as 3. Tables 4.3 and 4.4 show the average coherence score for all the topics extracted following the interactive algorithm compared to the other models. We can clearly see that in both cases, iLBLMA increases the topic quality indicated by the coherence score to a considerable extent. Especially in the case of LDA we see that the change is quite evident.

Table 4.3: Average coherence score of all topics for BBC news data

Model	K=10	K=15	K=20	K=25
iLBLMA	-1.08	-1.09	-1.30	-1.36
LBLMA	-1.17	-1.23	-1.40	-1.42
LBLA	-1.12	-1.19	-1.39	-1.47
LDA	-1.22	-1.40	-2.08	-1.87

Table 4.4: Average coherence score of all topics for 20 newsgroups data

Model	K=10	K=15	K=20	K=25
iLBLMA	-1.23	-1.35	-1.39	-1.50
LBLMA	-1.62	-1.37	-1.44	-1.53
LBLA	-1.42	-1.39	-1.58	-1.53
LDA	-1.69	-2.63	-1.90	-2.22

In addition to comparing the coherence scores, it would be a better experiment to check how the improvement in coherence score varies when the weights for topic proportions are changed. Figure 4.5 shows how the change

in weights impact the coherence score for BBC news data. We can see that when the weights for user defined probabilities was set as 0.8, the coherence increases much better. Even when the weights are 0.2, we can clearly see a slight improvement in topic quality. Though this is an obvious change it shows the flexibility of our model to serve the purpose even if the user does not have much knowledge about the subject. Similarly, figure 4.6 shows the

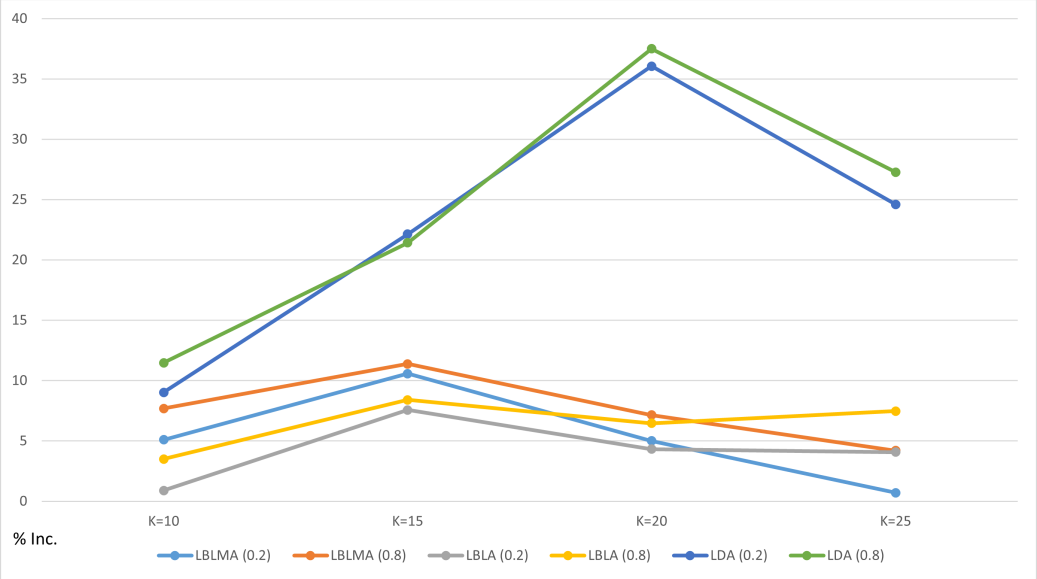

Figure 4.5: % Increase in topic quality for BBC news

variations when the weights are changed for the 20 newsgroups dataset. For this dataset, it was interesting to see that even though the coherence score increased, varying the weights did not have much impact on the topic quality. This is because even a slight bias to the topic probabilities by the user with the weight of 0.2 causes the coherence to reach the best possible value and did not require further modification. This effect is specific to this particular dataset. This shows the sturdiness of our model to perform at a constant level true to the LBLMA model when the user inputs do not have strong evidence. Overall, in all the cases, we find that our iLBLMA model has the capability to increase the quality of the topics extracted to a significant level.

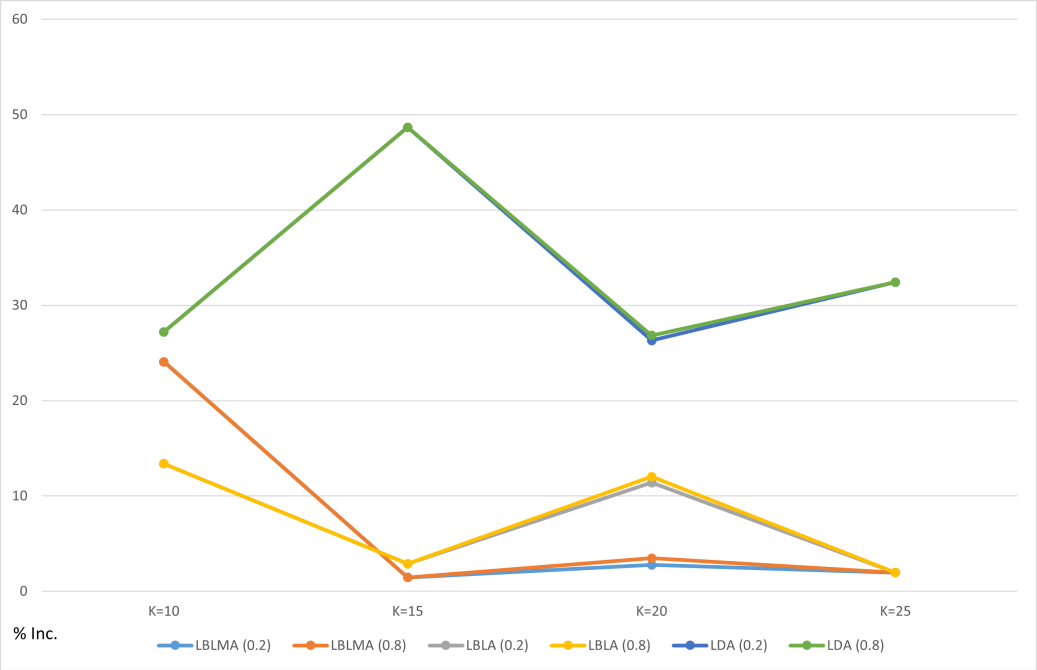

Figure 4.6: % Increase in topic quality for 20 newsgroups

Chapter 5

Biterm Learning for Recommendation Systems

Content based recommender systems play a vital role in applications related to user suggestions. In this chapter we explore ways to use our models to tackle the recommendation task. Recommendation systems have become an inseparable part of a variety of online services like web search, news articles, movies, etc. in recent years [79]. Most of the recent advancement in this field is centred on collaborative filtering [80] and content based filtering [79]. While the former method is based on modelling the activities of users with similar behaviour in a platform, the later works on modelling the likes of an individual user in the platform. Both approaches have their own merits and are used depending on the task at hand. LDA has also been used for creating recommendation systems [81, 82]. LDA can extract topics from the description of user activity which could help suggest new items that the user might be interested in.

It is well known that models which take into account, the co-occurrences of words, tend to give a boost for topic modelling tasks [65]. This made us to choose a design which incorporates the possibility of bigram words such as 'Thank you', 'high school', etc. This is relevant in our cases where most of our data uses short text descriptions. We will integrate this idea into our models to improve recommendations. We evaluate our model, with two challenging datasets. One of them is for anime recommendation and the other is for recommendation of movies from netflix. We estimate the performance of the model based on coherence score for both datasets. In addition, since we had

enough ground truth data to validate the netflix dataset, we estimate the accuracy of predictions as well.

The modifications required to convert our basic models is explained in Section 5.1. The experiments performed on the datasets with our models are detailed in Section 5.2.

5.1 Biterm Models

Contrary to bigrams where the probability of two words occurring together is considered, we take into account that logically these bigrams end up belonging to the same topic and consider them as bi-terms associated with the same topic as shown in Figure 5.1 and 5.2 resulting in biterm latent generalized Dirichlet mixture allocation (Bi-LGDMA) and biterm latent Beta-Liouville mixture allocation (Bi-LBLMA). This changes the equation for the topic word probability into,

$$p(w_{d(n-1)}, w_{dn} \mid z_{dn}, \vec{\beta}) = \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \beta_{kv}^{w_{d(n-1)(v-1)} + w_{dnv}} \Big)^{z_{dnk}} \tag{5.1}$$

$w_{d(n-1)} = v_{n-1}$ and $w_{dn} = v_n$ in the above equation incorporates the dependency of adjacent words to the topic latent variable.

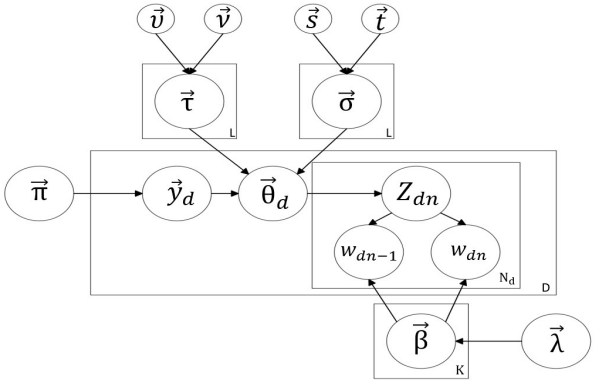

Figure 5.1: Graphical representation of Bi-LGDMA

The rest of the equations for both LGDMA and LBLMA remains the same and hence the variational solutions. Pertaining to these changes, the variational solutions of the hyper parameters δ_{dnk} and λ_{kv} is given by,

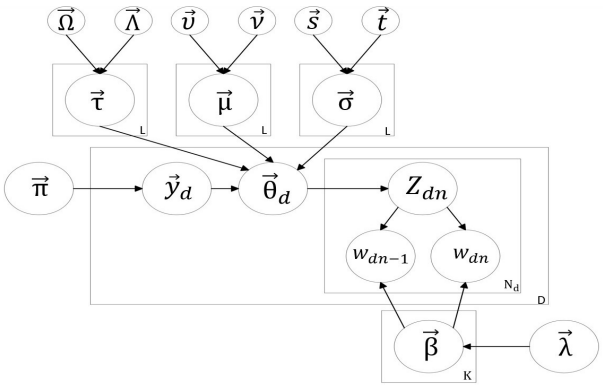

Figure 5.2: Graphical representation of Bi-LBLMA

$$\delta_{dnk} = \exp\Big(\big[w_{d(n-1)(v-1)} + w_{dnv}\big] \big\langle \ln \beta_{kv} \big\rangle + \big\langle \ln \theta_{dk} \big\rangle \Big) \tag{5.2}$$

$$\lambda^*_{kv} = \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} \phi_{dnk} \big[w_{d(n-1)v} + w_{dnv}\big] \tag{5.3}$$

5.2 Experimental results

To evaluate the performance of our model, we build a system for anime recommendation based on a dataset in Kaggle containing information about anime[1] and another for recommending movies based on data from netflix prize data [2]. We compare our model with widely used LDA and examine how our models weigh up against unmodified latent generalized Dirichlet allocation (LGDA) and latent Beta-Liouville allocation (LBLA) models. The idea of our recommendation system is that we find the Euclidean distance between the document topic proportions ϕ_{dk} of the query document and the rest of the documents. We can then find the top N recommendations for that query. The following subsections detail our experiments for the two datasets.

[1]https://www.kaggle.com/datasets/marlesson/myanimelist-dataset-animes-profiles-reviews

[2]https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

5.2.1 Anime Recommendation

This dataset consisted of 2 files containing information about anime, reviews of users and user profile details. The anime file had around 16K anime details like, title, synopsis, genre, airing date, etc. The profiles file had details of users and the anime they have added as favourites. The reviews file has information on the reviews the user has written for different anime. All these data has been extracted from https://myanimelist.net. From the anime details file, the data that helps for content based recommendation is mainly the synopsis. However, the synopsis was not available for some of the anime within the data. Hence we used the myanimelist API to extract missing synopsis. There were cases in which some of the titles refer to a parent anime and the description of parent anime was taken in these cases. We ignore anime where the synopsis is too short. After applying these constraints we were left with around 1126 anime to use for our content based recommendation system.

In the case of this dataset, there were a very few user profiles who had more than 20 anime in their favourites list which was not enough to evaluate our models. To understand the relevance of the topics that have been extracted by our model, we calculated UMass coherence score [60] explained in 2.7.2. Table 5.1 shows the coherence scores of topics derived from LDA, latent generalized Dirichlet allocation (LGDA), latent Beta-Liouville allocation (LBLA) and Bi-LGDMA and Bi-LBLMA for different values of L.

It can be seen that using a GD and BL prior helps in obtaining better topics with a higher coherence score. Bi-LBLMA performs better that Bi-LGDMA according to our experiments, which is due to the fact that choosing the parameters for Bi-LGDMA is a little harder than Bi-LBLMA. We calculated the coherence scores for different values of K to find the correct number of topics for the model. The best results were observed when K was set to 5. Figure 5.3 shows this much more clearly. Both Bi-LGDMA and Bi-LBLMA performed well when $L = 3$. In the case of Bi-LBLMA we see that the coherence is very close when $L = 3$ and $L = 4$. In these situations choosing the L as 3 or 4 will give similar recommendations.

This being a quantitative assessment of the model, to qualitatively see how the model performs, Table 5.2 shows few of the top ten suggestions for a query anime for the two models. 'Bleach' is an anime based on travelling between worlds through portals in the action genre. The anime suggested by Bi-LGDMA aligns with this concept of inter-dimensional portals and magic.

Table 5.1: Average coherence score of topics for Anime Data

Model	K=5	K=10	K=15	K=20	K=25
LDA	-1.67	-1.92	-2.13	-2.54	-2.86
LGDA	-1.48	-1.75	-1.95	-2.29	-2.56
Bi-LGDMA (L=2)	-1.37	-1.62	-1.83	-2.12	-2.21
Bi-LGDMA (L=3)	-1.32	-1.59	-1.79	-1.96	-2.08
Bi-LGDMA (L=4)	-1.36	-1.61	-1.85	-1.99	-2.10
Bi-LGDMA (L=5)	-1.35	-1.64	-1.85	-2.09	-2.17
LBLA	-1.42	-1.71	-2.07	-2.22	-2.25
Bi-LBLMA (L=2)	-1.35	-1.61	-1.81	-2.04	-2.11
Bi-LBLMA (L=3)	-1.28	-1.58	-1.76	-1.88	-2.10
Bi-LBLMA (L=4)	-1.28	-1.58	-1.80	-1.93	-2.09
Bi-LBLMA (L=2)	-1.31	-1.65	-1.79	-1.99	-2.11

Table 5.2: Query results for Anime data

S. No.	Bleach (Bi-LGDMA)	Dragon Ball (Bi-LBLMA)
1	Fullmetal Alchemist	Dragon Ball Z
2	Rosario to Vampire	Dragon Ball Super Movie: Broly
3	World Trigger	Boku no Hero Academia
4	FLCL	Yu-Gi-Oh Duel Monsters
5	Tenjou Tenge	Fate/stay night

Similarly, the test query for Bi-LBLMA was an anime called 'Dragon Ball' which involves super-human fighting. It is interesting to see that our model identified the sequel to the original anime followed by a few other anime like 'Boku no Hero Academia' which also falls under the same category.

5.2.2 Netflix movie recommendation

The Netflix dataset is bigger compared to the anime datset. The dataset consists of details pertaining to ratings fo different users for around 17000 movies released before the year 2006. However, the problem with this dataset

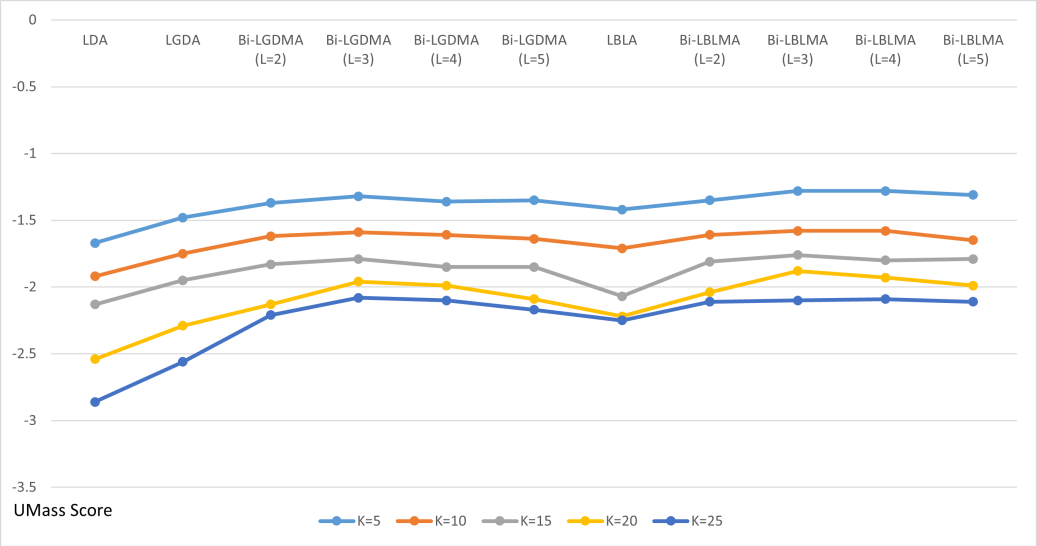

Figure 5.3: Coherence score for anime dataset for different values of K and L

is that the synopsis of movies were not available. Hence, we scraped the data from wikipedia pages to get this details and then used it for content based recommendation. We selected the movies released after 2000 so that we are aware of them to test qualitatively. This gave us around 4000 movies with description. From the user details, we consider that a user likes a movie when they rate it as 4 or 5. We selected users who had liked at least 300 movies. This left us with 900 users as ground truth. These conditions are only to quantitatively access our models and can be ignored in realtime applications. When queried with a movie that an user likes, if one of the top N recommendations by our model is present in the list of movies liked by that user, then we consider it as a hit. By using this logic, we can calculate the accuracy of our model by calculating the ratio of total number of hits to total number of queries.

We also calculate the coherence score of our topics as in the previous subsection which is shown in Table 5.3. We can see that both our models perform the best when $L = 2$ and $K = 5$. This is also seen in Figure 5.4. The performance improvement achieved by our models compared to the widely used LDA model proves the efficiency of our model to represent the topics better.

Table 5.3: Average coherence score of topics for Netflix Data

Model	K=5	K=10	K=15	K=20	K=25
LDA	-1.67	-1.92	-2.13	-2.54	-2.86
LGDA	-1.48	-1.75	-1.95	-2.29	-2.56
Bi-LGDMA (L=2)	-1.37	-1.62	-1.83	-2.12	-2.21
Bi-LGDMA (L=3)	-1.32	-1.59	-1.79	-1.96	-2.08
Bi-LGDMA (L=4)	-1.36	-1.61	-1.85	-1.99	-2.10
Bi-LGDMA (L=5)	-1.35	-1.64	-1.85	-2.09	-2.17
LBLA	-1.42	-1.71	-2.07	-2.22	-2.25
Bi-LBLMA (L=2)	-1.35	-1.61	-1.81	-2.04	-2.11
Bi-LBLMA (L=3)	-1.28	-1.58	-1.76	-1.88	-2.10
Bi-LBLMA (L=4)	-1.28	-1.58	-1.80	-1.93	-2.09
Bi-LBLMA (L=2)	-1.31	-1.65	-1.79	-1.99	-2.11

In addition to these analysis Table 5.4 shows the accuracy of different models. Though both Bi-LGDMA and Bi-LBLA give comparatively better accuracy for our model, the improvement for Bi-LGDMA is not that much when compared to Bi-LBLA. Similar to the last experiment, we also check

Table 5.4: Accuracy of recommendation at $N = 15$ for Netflix Data

Model	Accuracy
LDA	85.59
LGDA	84.40
Bi-LGDMA	86.00
LBLA	86.50
Bi-LBLMA	87.36

the quality of recommendations for two sample queries. This is shown in Table 5.5. We can see that Bi-LGDMA recommends a set of teenage and kids action movies like 'Agent Cody Banks' when queried with the movie 'THe Pacifier' which is a kids action comedy. In the case of Bi-LGDMA 'Resident Evil' is a zombie movie where the virus causes the people to attack the non-infected people. The recommendations from our model found similar

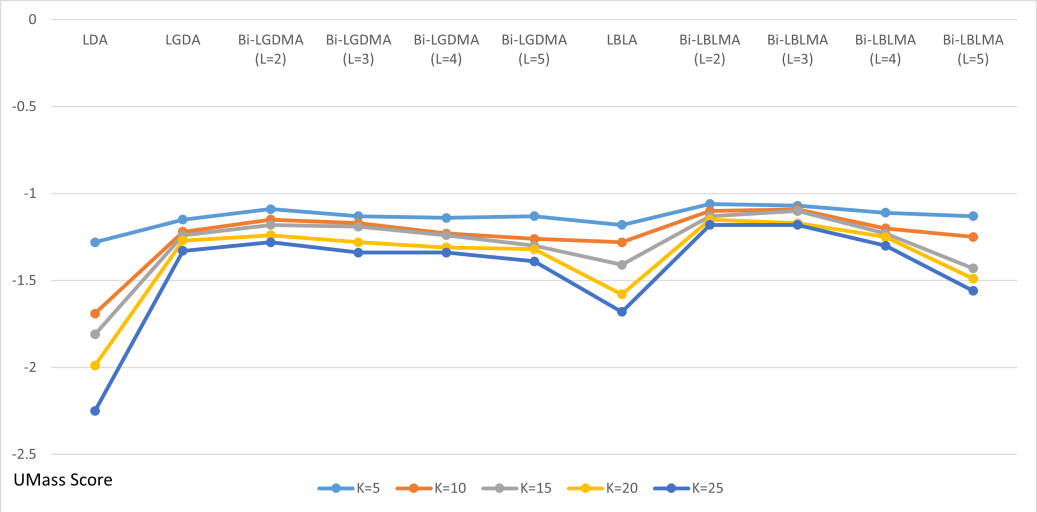

Figure 5.4: Coherence score for anime dataset for different values of K and L

plot lines like 'Dawn of the dead', 'Sasquatch', etc which are movies based on virus outbreak, hunted by animals and so on.

Table 5.5: Query results for Netflix data

S. No.	The Pacifier (Bi-LGDMA)	Resident Evil (Bi-LBLMA)
1	Agent Cody Banks	Dawn of the Dead
2	Agent Cody Banks 2: Destination London	Sasquatch
3	Lilo and Stitch 2	Wrong Turn
4	101 Dalmations II: Patch's London Adventure	Evil Remains
5	Mean Creek	Dead Birds

Chapter 6

Nonparametric Approach for Multilingual Data

In this chapter, we extend finite mixture model to infinite case to provide flexibility in modelling various topics to define proper number of clusters automatically [83]. It is notable that in the previous chapters we experimented with different values of L to find the optimal number of components [84]. Here we provide a solution to avoid this problem.

In the broad field of topic modelling, one of the main assumptions that could not be generalized is that the language of resources is English. Lots of models that have been conventionally used were designed to model monolingual contexts only and work with monolingual resources [2]. Considering the ongoing increase in technology specially in using online resources (for instance social media which connect various parts of the world together), more content from other languages besides English are becoming available. However, translating these valuable documents to English and using them in NLP algorithms that just work with one language is a great challenge and it is so costly and needs lots of time. Thus, there is a growing interest in finding solutions which could help scientists and industries to work with language-independent text mining tools without needing any translation resources.

To tackle this issue, multilingual NLP has been introduced and helped scientists to extract information regarding topics from various data sources and documents [85,86]. In this method, various languages are tied together which helps in discovering the connections in the languages of interest and building coherent topics across them. This helps us in indexing similar topics across

multiple languages which helps in multilingual document retrieval [87–91]. Also, we don't have any linguistic assumption about documents or data that we intend to model. This capability empowers our model to relax the constraint of modelling just one language and identifies similar patterns across multiple corpora in various languages. Such models with their power of inference on documents could be interesting in many applications [92–95].

We assume that our model has a nonparametric structure [96] which provides us considerable flexibility to model several topics in multiple languages. To do so, we use Dirichlet process (DP) [97, 98] and extend our finite mixture model to infinite case. This elegant method helps to address another task which is defining model complexity. Conventionally, some criteria such as Akaike information criterion [99], Bayes information criterion [100], minimum description length and minimum message length [34] have been applied to define proper number of clusters. But these methods are time consuming as we need to check them for various numbers of clusters. We evaluate the performance of our model with a real world dataset with two languages, English and French. We measure the quality of topics by comparing the coherence scores of the different models. We measure the similarity between topics in different languages with Jaccard index. Our experimental outcomes demonstrate the practicality of our proposed model in finding topics by processing multi-lingual documents. Though the equations and definitions might look repetitive, the presence of multiple languages causes slight differences in most of the equations. Hence we redefine all the equations in this chapter to avoid confusions.

In section 6.1, we explain in detail how to construct Dirichlet process based LGDA (DP-LGDA) and Dirichlet process base LBLA (DP-LBLA) models respectively. In section 6.2, we explain the learning method by proposing a variational framework. This is followed by Section 6.3 which is devoted to experimental results.

6.1 Model Description

In this section, we define the mathematical model for multilingual topic extraction with Dirichlet process mixture allocation with generalized Dirichlet and Beta-Liouville priors. First, we provide a general description of a topic model and then define the required forms for DP-LGDA and DP-LBLA.

Let us consider a set of D documents in M different languages, where,

$d = \{1, 2, ..., D\}$ and $m = \{1, 2, ..., M\}$ represent the d^{th} document and m^{th} language respectively. Each document d in language m can be represented as a word vector $\vec{w}_{md} = \left(w_{md1}, w_{md2}, ..., w_{mdN_{md}}\right)$, where, N_{md} is the number of words in that particular document. The n^{th} word in a document can be represented by an indicator vector which is V_m dimensional, corresponding to the vocabulary size of language m following the rule, $w_{mdnv} = 1$ when the word w_{mdn} is the same as the word v_m in the the vocabulary and 0 otherwise. Similarly, we also define a latent indicator variable $\mathcal{Z}_m = \{\vec{z}_{md}\} = \{\vec{z}_{mdn}\}$ showing which of the K topics the word belongs to based on the criteria $z_{mdnk} = 1$ if word w_{mdn} is present in topic k and 0 if not. Each language has a separate variable $\vec{\beta}_{mk}$ which describes the distribution of words in each topic, given by, $\vec{\beta}_{mk} = (\beta_{mk1}, \beta_{mk2}, ..., \beta_{mkV_m})$. To define the prior for the topic distribution for each document in a general manner, let's say $p(\vec{\theta} \mid \Phi)$ is the prior given the parameter of that distribution Φ. In the case of LDA this distribution is Dirichlet. It is to be noted that in our case the topic probabilities are drawn from an infinite mixture model. $\mathcal{Y} = (\vec{y}_1, \vec{y}_2, ..., \vec{y}_D)$ is the indicator matrix which stipulates which cluster the document belongs to, where $\vec{y}_d$ is L dimensional with $y_{dl} = 1$ when the document d belongs to cluster l. Here L is the truncation level set for the Dirichlet process mixture. $\mathcal{Y}$ is a multinomial distribution with parameters $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_L)$ corresponding to the mixing coefficient and follows the constraint $\sum_{l=1}^{L} \pi_l = 1$. The mixing coefficients here will follow a stick breaking approach to construct a DP model. The main idea here is to define a common set of topic proportion vectors $\vec{\theta}$ which is shared by all the languages. This restricts the topic proportion vectors that each document can take and forces the topic word proportions across multiple languages to have a similar structure. Doing this helps us to extract parallel topics across languages. The generative process for our multilingual model can be written as follows:

- For each language corpus m in the dataset:
  - For each word vector $\vec{w}_{md}$ in that corpus:
    - Draw component l from the mixture $y_d = l \backsim DirichletProcess(\vec{\pi})$
    - Draw topic proportions $\vec{\theta}_d \mid y_d = l$ from a mixture of L distributions
    - For each word n of the N_d words in document $\vec{w}_{md}$
      - Draw topic $z_{mdn} = k \backsim Multinomial(\vec{\theta}_d)$

· Draw word $w_{mdn} = v_m \mid z_{mdn} = k \backsim Multinomial(\vec{\beta}_{z_{mdn}})$

The marginal likelihood of this multilingual topic model can thus be written as,

$$p(W \mid \vec{\pi}, \vec{\Phi}, \vec{\beta}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \int \Bigg[\Big(\sum_{y_d} p(\vec{\theta}_d \mid y_d, \vec{\Phi}) p(y_d \mid \vec{\pi}) \Big) \times \prod_{n=1}^{N_d} \sum_{z_{mdn}} p(w_{mdn} \mid z_{mdn}, \vec{\beta_m}) p(z_{mdn} \mid \vec{\theta}_d) \Bigg] d\vec{\theta}_d \tag{6.1}$$

for a multilingual corpus W.

6.1.1 Dirichlet process based latent generalized Dirichlet allocation

As explained earlier, using a generalized Dirichlet prior helps overcome the drawbacks of Dirichlet distribution by providing a general covariance matrix [30]. When we use a generalized Dirichlet prior for the topic proportions $\vec{\theta}_d$, the distribution for each topic k takes the form,

$$p(\theta_{dk} \mid \sigma_{lk}, \tau_{lk}) = \frac{\Gamma(\tau_{lk} + \sigma_{lk})}{\Gamma(\tau_{lk})\Gamma(\sigma_{lk}))} \theta_{dk}^{\sigma_{lk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj} \right)^{\gamma_{lk}} \tag{6.2}$$

where $(\sigma_{l1}, \sigma_{l2}, ..., \sigma_{lN_d}, \tau_{l1}, \tau_{l2}, ..., \tau_{lN_d})$ are the parameters of GD distribution and $\gamma_k = \tau_k - \tau_{k+1} - \sigma_{k+1}$ for $k = 1, 2, ..., K-1$ and $\gamma_k = \sigma_k - 1$ for $k = K$. Since considering mixture of distributions help us to improve the topic model [101], we consider a mixture of GD distributions as prior for our model. Thus we can write the prior for our topic proportions as,

$$p(\vec{\theta}_d \mid \vec{y}_d, \vec{\sigma}, \vec{\tau}) = \prod_{l=1}^{\infty} \prod_{k=1}^{K} \Big(p(\theta_{dk} \mid \sigma_{lk}, \tau_{lk}) \Big)^{y_{dl}} \tag{6.3}$$

Since, $\vec{y}_d$ is a multinomial with parameter π, we can write $p(\vec{y}_d)$ as,

$$p(\vec{y}_d) = \prod_{l=1}^{\infty} \pi_l^{y_{dl}} \tag{6.4}$$

By using a stick-breaking reconstruction of DP, replacing π_j as a function of κ_j, the equation becomes,

$$p(\mathcal{Y} \mid \vec{\kappa}) = \prod_{d=1}^{D} \prod_{l=1}^{\infty} \left[\kappa_l \prod_{o=1}^{l-1} (1 - \kappa_o) \right]^{y_{dl}} \tag{6.5}$$

The first part of Eq. 6.1 can hence be written as,

$$p(\vec{\theta}_d \mid \vec{y}_d, \vec{\sigma}, \vec{\tau}) p(\vec{y}_d \mid \vec{\pi}) \Big) = \prod_{l=1}^{\infty} \prod_{k=1}^{K} \left[\Big(\kappa_l \prod_{o=1}^{l-1} (1 - \kappa_o) \Big) \Big(p(\theta_{dk} \mid \sigma_{lk}, \tau_{lk}) \Big) \right]^{y_{dl}} \tag{6.6}$$

$p(w_{mdn} \mid z_{mdn} \vec{\beta}_m)$ and $p(z_{mdn} \mid \vec{\theta}_d)$ are multinomials given by,

$$p(w_{mdn} \mid z_{mdn}, \vec{\beta}_m) = \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \beta_{mkv}^{w_{mdnv}} \Big)^{z_{mdnk}} \tag{6.7}$$

$$p(z_{mdn} \mid \vec{\theta}_d) = \prod_{k=1}^{K} \theta_{dk}^{z_{mdnk}} \tag{6.8}$$

We use Gamma priors which has proven to be an adequate alternative [29]. Hence the priors for the parameters of GD is given by,

$$p(\sigma_{lk}) = \mathcal{G}(\sigma_{lk} \mid \upsilon_{lk}, \nu_{lk}) = \frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})} \sigma_{lk}^{\upsilon_{lk}-1} e^{-\nu_{lk}\sigma_{lk}} \tag{6.9}$$

$$p(\tau_{lk}) = \mathcal{G}(\tau_{lk} \mid s_{lk}, t_{lk}) = \frac{t_{lk}^{s_{lk}}}{\Gamma(s_{lk})} \tau_{lk}^{s_{lk}-1} e^{-t_{lk}\tau_{lk}} \tag{6.10}$$

where $\mathcal{G}(\cdot)$ represents a Gamma distribution. The topic word proportions $\vec{\beta}_m$ with Dirichlet prior is given by,

$$p(\vec{\beta}_{mk} \mid \vec{\lambda}_{mk}) = \frac{\Gamma(\sum_{v=1}^{V_m} \lambda_{mkv})}{\prod_{v=1}^{V_m} \Gamma(\lambda_{mkv})} \prod_{v=1}^{V_m} \beta_{mkv}^{\lambda_{mkv}-1} \tag{6.11}$$

Assuming a variational prior for $\vec{\theta}_d$ helps us to simplify the inference process. Hence we define the equation,

$$p(\vec{\theta}_d \mid \vec{g}_d, \vec{h}_d) = \prod_{k=1}^{K} \frac{\Gamma(g_{dk} + h_{dk})}{\Gamma(g_{dk})\Gamma(h_{lk})} \theta_{dk}^{g_{dk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj} \right)^{\zeta_{dk}} \tag{6.12}$$

where, $\zeta_{dk} = h_{dk} - g_{d(k-1)} - h_{d(k-1)}$ while $k \leq K-1$ and $\zeta_{dk} = h_{dk} - 1$ when $k = K$. Similarly, we also place a Beta distribution to define $\vec{\kappa}$ with hyperparameters $\vec{\omega}$ which gives,

$$p(\vec{\kappa} \mid \vec{\omega}) = \prod_{l=1}^{\infty} Beta(1, \omega_l) = \prod_{l=1}^{\infty} \omega_l (1 - \kappa_l)^{\omega_l - 1} \tag{6.13}$$

Following the approach used in [102], we introduce Gamma priors to the stick lengths as,

$$p(\vec{\omega}) = \mathcal{G}(\vec{\omega} \mid \vec{a}, \vec{b}) = \prod_{l=1}^{\infty} \frac{b_l^{a_l}}{\Gamma(a_l)} \omega_j^{a_l - 1} e^{-b_l \omega_l} \tag{6.14}$$

Based on these equations, we can write the joint distribution of the posterior as,

$$\begin{aligned} p(W, \Theta) =& p(W \mid \mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\sigma}, \vec{\tau}, \mathcal{Y}) \\ =& p(\vec{W} \mid \mathcal{Z}, \vec{\beta}) p(\vec{z} \mid \vec{\theta}) p(\vec{\theta} \mid \vec{\sigma}, \vec{\tau}, \mathcal{Y}) p(\mathcal{Y} \mid \vec{\kappa}) p(\vec{\kappa} \mid \vec{\omega}) p(\vec{\omega}) \\ & p(\vec{\theta} \mid \vec{g}, \vec{h}) p(\vec{\beta} \mid \vec{\lambda}) p(\vec{\sigma} \mid \vec{v}, \vec{\nu}) p(\vec{\tau} \mid \vec{s}, \vec{t}) \end{aligned} \tag{6.15}$$

Given $\Theta = \{\mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\sigma}, \vec{\tau}, \mathcal{Y}\}$ which represents all the parameters in our model. We can represent our model as a plate diagram shown in Fig. 6.1.

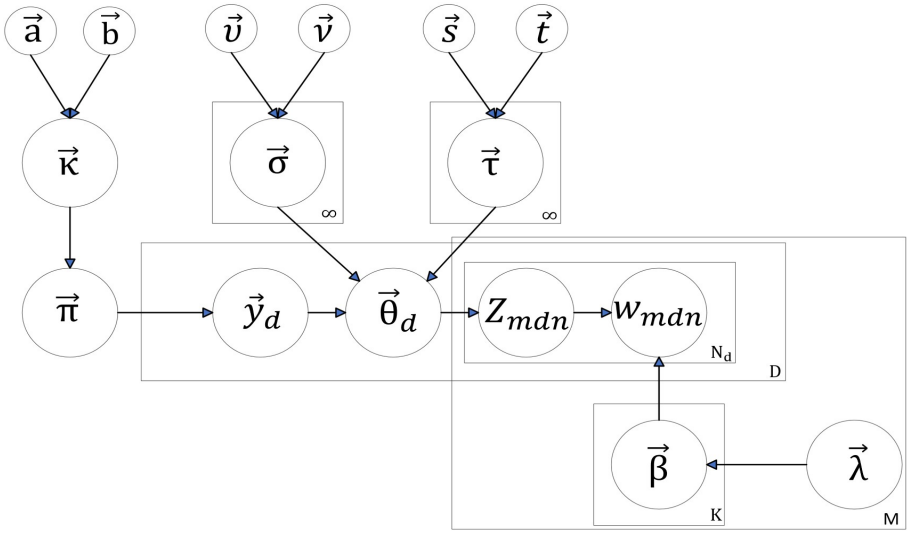

Figure 6.1: Plate model of DP-LGDA

6.1.2 Dirichlet process based latent Beta-Liouville mixture allocation

We can construct the DP-LBLA with the same definitions considered for DP-LGDA just by replacing the GD prior used in Eq. 6.2 with the BL distribution. The prior in this case can be written as,

$$p(\vec{\theta}_d \mid \vec{y}_d, \vec{\mu}, \vec{\sigma}, \vec{\tau}) = \prod_{l=1}^{L}\prod_{k=1}^{K}\left[\frac{\Gamma(\sum_{k=1}^{K}\mu_{lk})}{\prod_{k=1}^{K}\Gamma(\mu_{lk})}\frac{\Gamma(\sigma_l+\tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)}\theta_{dk}^{\mu_{lk}-1}\right. \\ \left.\times\Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{\sigma_l-\sum_{k=1}^{K}\mu_{lk}}\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{\tau_l-1}\right. \tag{6.16}$$

where $(\mu_{l1}, \mu_{l2}, ..., \mu_{lN_d}, \sigma_l, \tau_l)$ are the parameters of Beta-Liouville distribution. The Gamma priors for DP-LBLA can be similarly written as,

$$p(\mu_{lk}) = \mathcal{G}(\mu_{lk} \mid \upsilon_{lk}, \nu_{lk}) = \frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})}\mu_{lk}^{\upsilon_{lk}-1}e^{-\nu_{lk}\mu_{lk}} \tag{6.17}$$

$$p(\sigma_l) = \mathcal{G}(\sigma_l \mid s_l, t_l) = \frac{t_l^{s_l}}{\Gamma(s_l)}\sigma_l^{s_l-1}e^{-t_l\sigma_l} \tag{6.18}$$

$$p(\tau_l) = \mathcal{G}(\sigma_l \mid \Omega_l, \Lambda_l) = \frac{\Lambda_l^{\Omega_l}}{\Gamma(\Omega_l)}\tau_l^{\Omega_l-1}e^{-\Lambda_l\tau_l} \tag{6.19}$$

Changing the prior to BL distribution also changes the variational prior in Eq. 6.12 to,

$$p(\vec{\theta}_d \mid \vec{f}_d, g_d, h_d) = \prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}f_{dk})}{\prod_{k=1}^{K}\Gamma(f_{dk})}\frac{\Gamma(g_d+h_d)}{\Gamma(g_d)\Gamma(h_d)}\theta_{dk}^{f_{dk}-1} \\ \times\Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{g_d-\sum_{k=1}^{K}f_{dk}}\Big[1-\sum_{k=1}^{K}\theta_{dk}\Big]^{h_d-1} \tag{6.20}$$

Reflecting these changes, the joint likelihood can now be written with respect to the parameters Θ as,

$$p(W,\Theta) = p(W \mid \mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\mu}, \vec{\sigma}, \vec{\tau}, \mathcal{Y}) \tag{6.21}$$

$$= p(\vec{W} \mid \mathcal{Z}, \vec{\beta})p(\vec{z} \mid \vec{\theta})p(\vec{\theta} \mid \vec{\mu}, \vec{\sigma}, \vec{\tau}, \mathcal{Y})p(\mathcal{Y} \mid \vec{\kappa})p(\vec{\kappa} \mid \vec{\omega})p(\vec{\omega}) \\ p(\vec{\theta} \mid \vec{f}, \vec{g}, \vec{h})p(\vec{\beta} \mid \vec{\lambda})p(\vec{\mu} \mid \vec{v}, \vec{\nu})p(\vec{\sigma} \mid \vec{s}, \vec{t})p(\vec{\tau} \mid \vec{\Omega}, \vec{\Lambda}) \tag{6.22}$$

The plate model of DP-LBLA is shown in Fig. 6.2

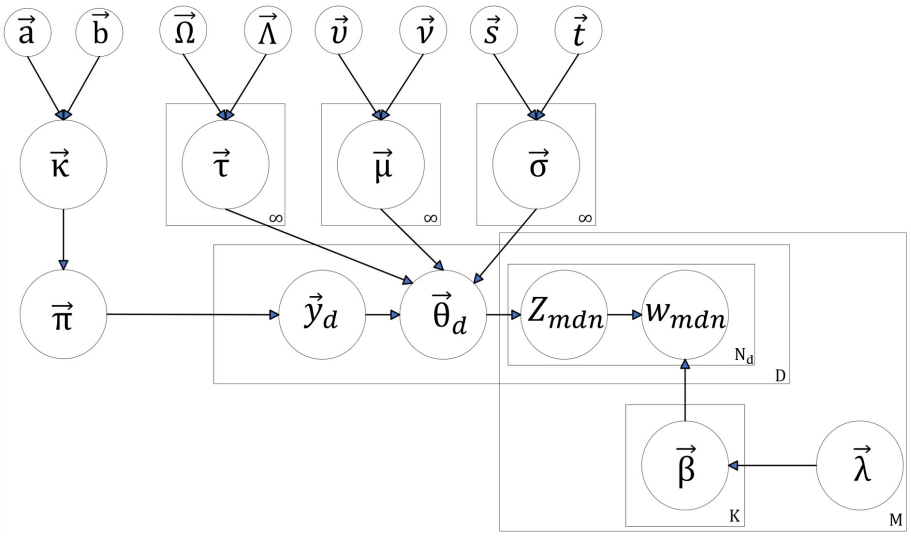

Figure 6.2: Plate model of DP-LBLA

6.2 Variational Solutions

Calculating the variational solutions as described earlier in section 3.2 for Eq. 6.15 results in the following equations:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D}\prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K} \phi_{mdnk}^{z_{mdnk}}, Q(\vec{\kappa}) = \prod_{l=1}^{L} Beta\Big(\kappa_l \mid c_l^*, d_l^*\Big) \tag{6.23}$$

$$Q(\vec{\sigma}) = \prod_{l=1}^{L}\prod_{k=1}^{K} \frac{{\nu_{lk}^*}^{\upsilon_{lk}^*}}{\Gamma(\upsilon_{lk}^*)} \sigma_{lk}^{\upsilon_{lk}^*-1} e^{-\nu_{lk}^*\sigma_{lk}}, Q(\vec{\tau}) = \prod_{l=1}^{L}\prod_{k=1}^{K} \frac{{t_{lk}^*}^{s_{lk}^*}}{\Gamma(s_{lk}^*)} \tau_{lk}^{s_{lk}^*-1} e^{-t_{lk}^*\tau_{lk}} \tag{6.24}$$

$$Q(\vec{\beta}) = \prod_{k=1}^{K}\prod_{v=1}^{V_m} \frac{\Gamma(\sum_{v=1}^{V_m}\lambda_{kv}^*)}{\prod_{v=1}^{V_m}\Gamma(\lambda_{kv}^*)} \beta_{kv}^{\lambda_{kv}^*-1} \tag{6.25}$$

$$Q(\vec{\theta}) = \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(g_{dk}^* + h_{dk}^*)}{\Gamma(g_{dk}^*)\Gamma(h_{lk}^*))} \theta_{dk}^{g_{dk}^*-1} \left(1 - \sum_{j=1}^{k}\theta_{dj}\right)^{\zeta_{dk}^*} \tag{6.26}$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{mdnk} = \frac{\delta_{mdnk}}{\sum_{k=1}^{K} \delta_{mdnk}} \tag{6.27}$$

$$\rho_{dl} = \exp\left\{ \left\langle \ln \kappa_l \right\rangle + \sum_{s=1}^{l-1} \left\langle \ln(1-\kappa_s) \right\rangle + \mathcal{R}_l + \sum_{k=1}^{K} (\sigma_{lk}-1) \left\langle \ln \theta_{dk} \right\rangle + \gamma_{lk} \left\langle 1 - \sum_{j=1}^{k} \theta_{dj} \right\rangle \right\} \tag{6.28}$$

$$\delta_{mdnk} = \exp(\left\langle \ln \beta_{mkv} \right\rangle + \left\langle \ln \theta_{dk} \right\rangle) \tag{6.29}$$

Here, $\vec{\mathcal{R}}$ is the taylor series approximations of $\left\langle \ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \right\rangle$ and is given by,

$$\begin{aligned}
\vec{\mathcal{R}} =& \ln \frac{\Gamma(\overline{\sigma}+\overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\sigma})\big](\left\langle \ln \sigma \right\rangle - \ln \overline{\sigma}) \\
&+ \overline{\tau}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\tau})\big](\left\langle \ln \tau \right\rangle - \ln \overline{\tau}) \\
&+ 0.5\overline{\sigma}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\sigma})\big]\left\langle (\ln \sigma - \ln \overline{\sigma})^2 \right\rangle \\
&+ 0.5\overline{\tau}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\tau})\big]\left\langle (\ln \tau - \ln \overline{\tau})^2 \right\rangle \\
&+ \overline{\sigma}\,\overline{\tau}\,\Psi'(\overline{\sigma}+\overline{\tau})(\left\langle \ln \sigma \right\rangle - \ln \overline{\sigma})(\left\langle \ln \tau \right\rangle - \ln \overline{\tau})
\end{aligned} \tag{6.30}$$

$$\begin{aligned}
\upsilon^*_{lk} =& \upsilon_{lk} + \sum_{d=1}^{D} \left\langle y_{dl} \right\rangle \Bigg[\Psi\big(\overline{\sigma}_{lk} + \overline{\tau}_{lk}\big) - \Psi\big(\overline{\sigma}_{lk}\big) \\
&+ \overline{\tau}_{lk} \Psi'\big(\overline{\sigma}_{lk} + \overline{\tau}_{lk}\big)\big(\left\langle \ln \tau_{lk} \right\rangle - \ln \overline{\tau}_{lk}\big) \Bigg] \overline{\sigma}_{lk}
\end{aligned} \tag{6.31}$$

$$\begin{aligned}
s^*_{lk} =& s_{lk} + \sum_{d=1}^{D} \left\langle y_{dl} \right\rangle \Bigg[\Psi\big(\overline{\tau}_{lk} + \overline{\sigma}_{lk}\big) - \Psi\big(\overline{\tau}_{lk}\big) \\
&+ \overline{\sigma}_{lk} \Psi'\big(\overline{\tau}_{lk} + \overline{\sigma}_{lk}\big)\big(\left\langle \ln \sigma_{lk} \right\rangle - \ln \overline{\sigma}_{lk}\big) \Bigg] \overline{\tau}_{lk}
\end{aligned} \tag{6.32}$$

$$\nu^*_{lk} = \nu_{lk} - \sum_{d=1}^{D} \left\langle y_{dl} \right\rangle \left\langle \ln \theta_{dk} \right\rangle \tag{6.33}$$

$$t^*_{lk} = t_{lk} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \left\langle \ln \Big[1 - \sum_{j=1}^{K} \theta_{dj} \Big] \right\rangle \tag{6.34}$$

$$g^*_{dk} = g_{dk} + \sum_{n=1}^{N_d} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \sigma_{lk} \tag{6.35}$$

$$h^*_{dk} = h_{dk} + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \tau_{lk} + \sum_{kk=k+1}^{K} \phi_{dn(kk)} \tag{6.36}$$

$$\lambda^*_{mkv} = \lambda_{mkv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V_m} \phi_{mdnk} w_{mdnv} \tag{6.37}$$

$$c^*_l = 1 + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle, d^*_l = \big\langle \kappa_j \big\rangle + \sum_{d=1}^{D} \sum_{s=l+1}^{L} \big\langle y_{ds} \big\rangle \tag{6.38}$$

$$a^*_l = a_l + 1, b^*_l = b_l - \big\langle ln(1 - \kappa_l) \big\rangle \tag{6.39}$$

In the above equations, $\big\langle \cdot \big\rangle$ indicates expectation of the variable and $(\bar{\cdot})$ is the mean of the variable. The values of these expectations [29] and mean are given by,

$$\big\langle \ln \theta_{dk} \big\rangle = \sum_{j=1}^{k} \big(\Psi(g_{dk}) - \Psi(g_{dk} + h_{dk}) \big) \tag{6.40}$$

$$\Big\langle 1 - \sum_{j=1}^{k} \theta_{dj} \Big\rangle = \sum_{j=1}^{k} \big(\Psi(h_{dk}) - \Psi(g_{dk} + h_{dk}) \big) \tag{6.41}$$

$$\overline{\sigma}_{lk} = \frac{\upsilon^*_{lk}}{\nu^*_{lk}}, \big\langle \ln \sigma_{lk} \big\rangle = \Psi\big(\upsilon^*_{lk}\big) - \ln \nu^*_{lk} \tag{6.42}$$

$$\big\langle \big(\ln \sigma_{lk} - \ln \overline{\sigma}_{lk} \big)^2 \big\rangle = \big[\Psi\big(\upsilon^*_{lk}\big) - \ln \upsilon^*_{lk} \big]^2 + \Psi'\big(\upsilon^*_{lk}\big) \tag{6.43}$$

$$\overline{\tau}_{lk} = \frac{s^*_{lk}}{t^*_{lk}}, \big\langle \ln \tau_{lk} \big\rangle = \Psi\big(s^*_{lk}\big) - \ln t^*_{lk} \tag{6.44}$$

$$\big\langle \big(\ln \tau_{lk} - \ln \overline{\tau}_{lk} \big)^2 \big\rangle = \big[\Psi\big(s^*_{lk}\big) - \ln s^*_{lk} \big]^2 + \Psi'\big(s^*_{lk}\big) \tag{6.45}$$

$$\left\langle z_{mdnk} \right\rangle = \phi_{mdnk}, \left\langle y_{dl} \right\rangle = r_{dl}, \left\langle \ln \beta_{mkv} \right\rangle = \Psi(\lambda_{mkv}) - \Psi(\sum_{f=1}^{V_m} \lambda_{mkf}) \quad (6.46)$$

$$\left\langle ln\kappa_l \right\rangle = \psi(c_l^*) - \psi(c_l^* + d_l^*), \left\langle ln(1-\kappa_l) \right\rangle = \psi(d_l^*) - \psi(c_l^* + d_l^*) \quad (6.47)$$

$\Psi(\cdot)$ and $\Psi(\cdot)'$ in the above equations indicate the digamma and trigamma functions respectively. We calculate equations 6.23 - 6.26 iteratively until convergence is achieved to find the optimal solutions.

6.2.1 Variational solutions for DP-LBLA

The variational solutions for Eq. 6.21 is more or less the same as in the previous section, except that some definitions of variables are different in addition to the obvious change in $Q(\vec{\theta})$. The variational solutions are:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D}\prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K} \phi_{dnk}^{z_{dnk}}, Q(\vec{\kappa}) = \prod_{l=1}^{L} Beta\Big(\kappa_l \mid c_l^*, d_l^*\Big) \quad (6.48)$$

$$Q(\vec{\mu}) = \prod_{l=1}^{L}\prod_{k=1}^{K} \frac{\nu_{lk}^{*\,\upsilon_{lk}^*}}{\Gamma(\upsilon_{lk}^*)} \mu_{lk}^{\upsilon_{lk}^*-1} e^{-\nu_{lk}^*\mu_{lk}}, Q(\sigma_l) = \prod_{l=1}^{L} \frac{t_l^{*s_l^*}}{\Gamma(s_l^*)} \sigma_l^{s_l^*-1} e^{-t_l^*\sigma_l} \quad (6.49)$$

$$Q(\tau_l) = \prod_{l=1}^{L} \frac{\Lambda_l^{*\Omega_l^*}}{\Gamma(\Omega_l^*)} \tau_l^{\Omega_l^*-1} e^{-\Lambda_l^*\tau_l}, Q(\vec{\beta}) = \prod_{k=1}^{K}\prod_{v=1}^{V} \frac{\Gamma(\sum_{v=1}^{V}\lambda_{kv}^*)}{\prod_{v=1}^{V}\Gamma(\lambda_{kv}^*)} \beta_{kv}^{\lambda_{kv}^*-1} \quad (6.50)$$

$$\begin{aligned} Q(\vec{\theta}) = & \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} f_{dk}^*)}{\Gamma(f_{dk}^*)} \frac{\Gamma(g_d^* + h_d^*)}{\Gamma(g_d^*)\Gamma(h_d^*)} \theta_{dk}^{f_{dk}^*-1} \\ & \times \Big[\sum_{k=1}^{K} \theta_{dk}\Big]^{g_d^* - \sum_{k=1}^{K} f_{dk}^*} \Big[1 - \sum_{k=1}^{K} \theta_{dk}\Big]^{h_d^*-1} \end{aligned} \quad (6.51)$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K} \delta_{dnk}} \tag{6.52}$$

$$\begin{aligned}\rho_{dl} = \exp\Bigg\{ & \Big\langle \ln \kappa_l \Big\rangle + \sum_{s=1}^{l-1} \Big\langle \ln(1-\kappa_s) \Big\rangle + \mathcal{R}_l + \mathcal{S}_l + (\mu_{lk}-1)\Big\langle \ln \theta_{dk} \Big\rangle \\ & + \Big(\sigma_l - \sum_{k=1}^{K} \mu_{lk}\Big)\Big\langle \ln \big[\sum_{k=1}^{K} \theta_{dk}\big] \Big\rangle + (\tau_l - 1)\Big\langle \ln \Big[1 - \sum_{k=1}^{K} \theta_{dk}\Big] \Big\rangle \Bigg\}\end{aligned} \tag{6.53}$$

Due to intractability, we use taylor series expansions for $\big\langle \frac{\Gamma(\sum_{k=1}^{K} \sigma_{lk})}{\Gamma(\sigma_{lk})} \big\rangle$ and $\big\langle \ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \big\rangle$ denoted by $\mathcal{R}$ and $\mathcal{S}$ respectively. The approximations are given as,

$$\begin{aligned}\mathcal{R}_l = & \ln \frac{\Gamma(\sum_{k=1}^{K} \mu_{lk})}{\prod_{k=1}^{K} \Gamma(\mu_{lk})} + \sum_{k=1}^{K} \overline{\mu}_{lk} \Big[\Psi\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk})\Big]\big[\big\langle \ln \mu_{lk} \big\rangle - \ln \overline{\mu}_{lk}\big] \\ & + \frac{1}{2} \sum_{k=1}^{K} \overline{\mu}_{lk}^2 \Big[\Psi'\big(\sum_{k=1}^{K} \overline{\mu}_{lk}\big) - \Psi'(\overline{\mu}_{lk})\Big] - \big\langle (\ln \mu_{lk} - \ln \overline{\mu}_{lk})^2 \big\rangle \\ & + \frac{1}{2} \sum_{a=1}^{K} \sum_{b=1, a\neq b}^{K} \overline{\mu}_{la} \overline{\mu}_{lb} \Big[\Psi'\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big)\big(\big\langle \ln \mu_{la} \big\rangle - \ln \overline{\mu}_{la}\big)\big(\big\langle \ln \mu_{lb} \big\rangle - \ln \overline{\mu}_{lb}\big)\Big]\end{aligned} \tag{6.54}$$

$$\begin{aligned}\vec{\mathcal{S}} = & \ln \frac{\Gamma(\overline{\sigma}+\overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\sigma})\big](\big\langle \ln \sigma \big\rangle - \ln \overline{\sigma}) \\ & + \overline{\tau}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\tau})\big](\big\langle \ln \tau \big\rangle - \ln \overline{\tau}) \\ & + 0.5\overline{\sigma}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\sigma})\big]\big\langle (\ln \sigma - \ln \overline{\sigma})^2 \big\rangle \\ & + 0.5\overline{\tau}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\tau})\big]\big\langle (\ln \tau - \ln \overline{\tau})^2 \big\rangle \\ & + \overline{\sigma}\,\overline{\tau}\,\Psi'(\overline{\sigma}+\overline{\tau})(\big\langle \ln \sigma \big\rangle - \ln \overline{\sigma})(\big\langle \ln \tau \big\rangle - \ln \overline{\tau})\end{aligned} \tag{6.55}$$

$$\begin{aligned}\upsilon_{lk}^{*} = & \upsilon_{lk} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \overline{\mu}_{lk} \Big[\Psi\Big(\sum_{k=1}^{K} \overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk}) \\ & + \Psi\Big(\sum_{k=1}^{K}\Big) \sum_{a\neq k}^{K} \big(\big\langle \ln \mu_{la} \big\rangle - \ln \overline{\mu}_{la}\big)\overline{\mu}_{la}\Big]\end{aligned} \tag{6.56}$$

$$\nu_{lk}^{*} = \nu_{lk} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\big\langle \ln \theta_{dk} \big\rangle - \Big\langle \ln \sum_{k=1}^{K} \theta_{dk} \Big\rangle \Big] \tag{6.57}$$

$$s_{l}^{*} = s_{l} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\Psi\big(\overline{\sigma}_{l} + \overline{\tau}_{l}\big) - \Psi\big(\overline{\sigma}_{l}\big) + \overline{\tau}_{l} \Psi'\big(\overline{\sigma}_{l} + \overline{\tau}_{l}\big) \big(\big\langle \ln \tau_{l} \big\rangle - \ln \overline{\tau}_{l} \big) \Big] \overline{\sigma}_{l} \tag{6.58}$$

$$t_{l}^{*} = t_{l} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big\langle \ln \Big[\sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{6.59}$$

$$\Omega_{l}^{*} = \Omega_{lk} + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big[\Psi\big(\overline{\tau}_{l} + \overline{\sigma}_{l}\big) - \Psi\big(\overline{\tau}_{l}\big) + \overline{\sigma}_{l} \Psi'\big(\overline{\tau}_{l} + \overline{\sigma}_{l}\big) \big(\big\langle \ln \sigma_{l} \big\rangle - \ln \overline{\sigma}_{l} \big) \Big] \overline{\tau}_{l}$$

$$\Lambda_{l}^{*} = \Lambda_{l} - \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle \Big\langle \ln \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{6.60}$$

$$f_{dk}^{*} = f_{dk} + \sum_{n=1}^{N_d} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \mu_{lk} \tag{6.61}$$

$$g_{d}^{*} = g_{d} + \sum_{n=1}^{N_d} \sum_{k=1}^{K} \big\langle z_{dnk} \big\rangle + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \sigma_{l} \tag{6.62}$$

$$h_{d}^{*} = h_{d} + \sum_{l=1}^{L} \big\langle y_{dl} \big\rangle \tau_{l} \tag{6.63}$$

$$c_{l}^{*} = 1 + \sum_{d=1}^{D} \big\langle y_{dl} \big\rangle, d_{l}^{*} = \big\langle \kappa_{j} \big\rangle + \sum_{d=1}^{D} \sum_{s=l+1}^{L} \big\langle y_{ds} \big\rangle \tag{6.64}$$

$$a_{l}^{*} = a_{l} + 1, b_{l}^{*} = b_{l} - \big\langle ln(1 - \kappa_{l}) \big\rangle \tag{6.65}$$

The expectations in these equations are defined with respect to BL distribution as follows:

$$\big\langle \ln \theta_{dk} \big\rangle = \Psi(f_{dk}) - \Psi\Big(\sum_{k=1}^{K} f_{dk}\Big) + \Psi(g_d) - \Psi(g_d + h_d) \tag{6.66}$$

$$\Big\langle \sum_{k=1}^{k} \theta_{dk} \Big\rangle = \sum_{k=1}^{k} \big(\Psi(g_d) - \Psi(g_d + h_d)\big) \tag{6.67}$$

$$\Big\langle 1 - \sum_{k=1}^{k} \theta_{dk} \Big\rangle = \sum_{k=1}^{k} \big(\Psi(h_d) - \Psi(g_d + h_d)\big) \tag{6.68}$$

$$\overline{\sigma}_{lk} = \frac{\upsilon^*_{lk}}{\nu^*_{lk}}, \big\langle \ln \sigma_{lk} \big\rangle = \Psi\big(\upsilon^*_{lk}\big) - \ln \nu^*_{lk} \tag{6.69}$$

$$\big\langle \big(\ln \sigma_{lk} - \ln \overline{\sigma}_{lk} \big)^2 \big\rangle = \big[\Psi\big(\upsilon^*_{lk}\big) - \ln \upsilon^*_{lk}\big]^2 + \Psi'\big(\upsilon^*_{lk}\big) \tag{6.70}$$

$$\overline{\sigma}_{l} = \frac{s^*_{l}}{t^*_{l}}, \big\langle \ln \sigma_{l} \big\rangle = \Psi\big(s^*_{l}\big) - \ln t^*_{l} \tag{6.71}$$

$$\big\langle \big(\ln \sigma_{l} - \ln \overline{\sigma}_{l} \big)^2 \big\rangle = \big[\Psi\big(s^*_{l}\big) - \ln s^*_{l}\big]^2 + \Psi'\big(s^*_{l}\big) \tag{6.72}$$

$$\overline{\tau}_{lk} = \frac{\Omega^*_{l}}{\Lambda^*_{l}}, \big\langle \ln \tau_{l} \big\rangle = \Psi\big(\Omega^*_{l}\big) - \ln \Lambda^*_{l} \tag{6.73}$$

$$\big\langle \big(\ln \tau_{l} - \ln \overline{\tau}_{l} \big)^2 \big\rangle = \big[\Psi\big(\Omega^*_{l}\big) - \ln \Omega^*_{l}\big]^2 + \Psi'\big(\Omega^*_{l}\big) \tag{6.74}$$

$$\big\langle z_{dnk} \big\rangle = \phi_{dnk}, \big\langle y_{dl} \big\rangle = r_{dl}, \big\langle \ln \beta_{kv} \big\rangle = \Psi(\kappa_{kv}) - \Psi(\sum_{f=1}^{V} \kappa_{kf}) \tag{6.75}$$

$$\big\langle \ln \kappa_l \big\rangle = \psi(c^*_l) - \psi(c^*_l + d^*_l), \big\langle \ln(1 - \kappa_l) \big\rangle = \psi(d^*_l) - \psi(c^*_l + d^*_l) \tag{6.76}$$

We follow the same process as before and compute equations 6.48 - 6.51 repeatedly until convergence.

6.3 Experimental Results

In order to evaluate our multi-lingual model, we choose a dataset which comprises transcripts of TED talks on varied topics from Kaggle [1]. The parallel dataset consists of talks from various disciplines like physics, environment, politics, relationships, pollution, space, etc. In order to perform a deep analysis pertaining to the quality of the extracted topics we keep things simple by choosing 99 talks which comprises of around 30 transcripts of talks closely related to three different topics namely, astrophysics, relationships and climate change. These talks do not exactly belong to the same class and might be slightly different in many cases. For example, a talk from astrophysics might be about space travel and aliens or experiments on dark matter. The variance in these topics with a small dataset helps us to see how our model is able to perform in situations where only limited data is available for learning. We calculate UMass coherence score [60] as in section 2.7.2 to evaluate the model

The first experiment we conducted is to test the quality of topics extracted by our models compared to other standard models. We use transcripts from only French and English to simplify analysis, however, the models will perform equally if compared with more languages as well. LDA being the basic and widely used topic extraction model will be our benchmark to compare, followed by Poly-LDA [103] which is a multilingual model based on LDA. We compare the coherence scores of the extracted topics with respect to each language separately. We also wanted to test how the model performs if a Dirichlet process mixture is not used for the mixture. In this case the parameter π_l will act as mixing coefficients of the model and the equations will transform accordingly. These models will be represented as 'Mix-LGDA' and 'Mix-LBLA' respectively. To study the effect of modifying the prior distributions, we also compare with 'DP-LDA' and 'mix-LDA' which are the LDA counterparts of our models.

Tables 6.1 and 6.2 show the coherence score for the different models for English and French languages respectively while varying the number of topics K. We can see that the coherence score is the highest when the number of topics is set as 5. Even though our data consisted of documents from three main categories, most of them had multidisciplinary concepts which was clearly captured by almost all of the models. LDA falters here when

[1] https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset

extracting topics in English in addition to mix-LBLA.

Table 6.1: Average coherence score of topics for TED talks transcripts in English

Model	K=3	K=5	K=7	K=9
LDA	-0.44	-0.45	-0.53	-0.61
Poly-LDA	-0.44	-0.43	-0.52	-0.59
Mix-LDA	-0.45	-0.44	-0.52	-0.52
DP-LDA	-0.42	-0.41	-0.48	-0.54
Mix-LGDA	-0.47	-0.43	-0.48	-0.53
DP-LGDA	-0.46	-0.43	-0.45	-0.51
Mix-LBLA	-0.39	-0.38	-0.44	-0.47
DP-LBLA	-0.39	-0.35	-0.41	-0.44

Table 6.2: Average coherence score of topics for TED talks transcripts in French

Model	K=3	K=5	K=7	K=9
LDA	-5.97	-5.29	-6.29	-6.84
Poly-LDA	-5.67	-5.67	-6.59	-7.02
Mix-LDA	-5.45	-5.25	-5.73	-6.88
DP-LDA	-5.33	-5.16	-5.75	-6.57
Mix-LGDA	-5.59	-5.05	-5.56	-6.88
DP-LGDA	-5.36	-4.90	-5.33	-5.88
Mix-LBLA	-5.27	-4.79	-5.43	-5.99
DP-LBLA	-5.10	-4.50	-5.25	-5.25

In general, we can observe that the coherence score is higher for mixture models without Dirichlet process assumption compared to LDA and poly-LDA. Similarly, with the Dirichlet process assumption, the models perform to the best compared to the rest. This pattern is clearly observed in Figs. 6.3 and 6.4 respectively. Though the coherence scores are different in scale for French and English, both languages follow a similar pattern.

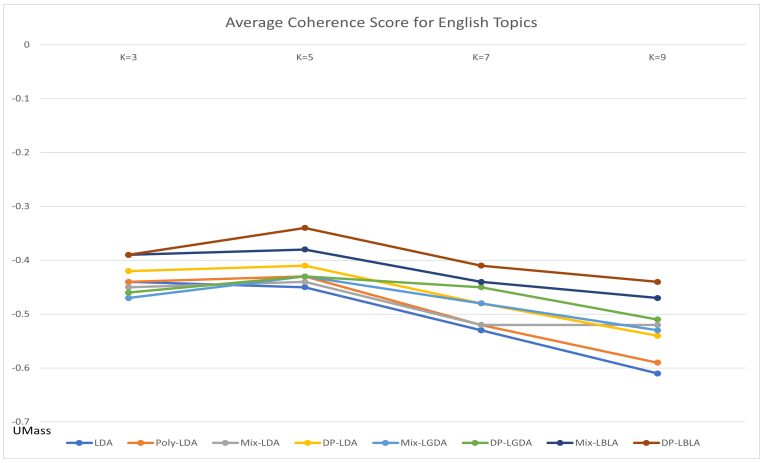

Figure 6.3: Coherence score for different models for varying number of topics in English

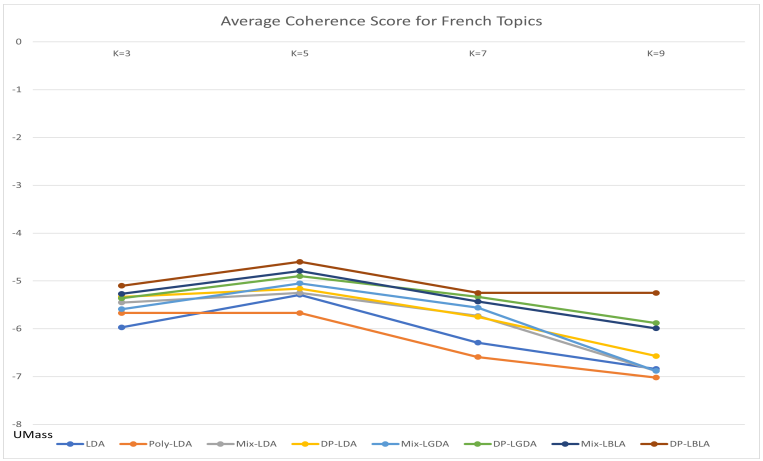

Figure 6.4: Coherence score for different models for varying number of topics in French

In order to analyse deeper, let us consider the topics extracted by our baseline poly-LDA and the best performing model in terms of coherence, DP-LBLA. Tables 6.3 and 6.4 show the French and English topics extracted by poly-LDA and DP-LBLA respectively. Looking at the topics we can see how good DP-LBLA is able to extract parallel topics. Topic 1 represents astrophysics, topic 2 is a set of common words in all topics, topic 3 is an

intersection between science and energy, topic 4 shows words corresponding to relationships and topic 5 is related to climate change. Interestingly, the word 'like' is present in almost all the languages as a major word. This is because it was used in almost all the talks by different people when they pause and give examples.

To check how similar the extracted topics are to each other, we calculate the Jaccard index between the set of words in each topic in both languages. F and E represents the set of words in the French topic and English topic respectively. The Jaccard index is then calculated with the formula:

$$Jac(F, E) = \frac{F \cap E}{F \cup E} \tag{6.77}$$

Table 6.3: Jaccard Index between English and French topics extracted by DP-LBLA

French	English	F∩E	F∪E	Jac
'lumi', 'donc', 'cette', 'toiles', 'espace', 'mati', 'galaxie', 'particules', 'galaxies', 'univers'	'like', 'dark', 'matter', 'black', 'light', 'universe', 'galaxy', 'galaxies', 'space', 'stars',	7	13	0.54
'quand', 'chose', 'cette', 'cela', 'bien', 'parce', 'comme', 'donc', 'alors', 'tout'	'find', 'mars', 'well', 'time', 'like', 'actually', 'think', 'life', 'going', 'know',	2	18	0.11
'soleil', 'surface','cette', 'solaire', 'comme', 'milliards', 'atmosph', 'syst', 'terre', 'plan'	'solar', 'planet', 'water', 'ocean', 'surface', 'energy', 'atmosphere', 'years', 'system', 'earth',	6	14	0.43
'tout', 'gens', 'cette', 'autre', 'cela', 'amour', 'quelqu', 'personne', 'rires', 'quand'	'want', 'brain', 'feel', 'person', 'really', 'think', 'laughter', 'love', 'like', 'people'	4	16	0.25
'donc', 'gens', 'pays', 'probl', 'climatique', 'cette', 'changement', 'tout', 'monde', 'cela',	'really', 'much', 'think', 'going', 'need', 'climate', 'global', 'change', 'world', 'people'	5	16	0.31

Table 6.4: Jaccard Index between English and French topics extracted by Poly-LDA

French	English	F∩E	F∪E	Jac
‘mati’, ‘lumi’, ‘plastique’, ‘espace’, ‘particules’, ‘toiles’, ‘galaxies’, ‘donc’, ‘cette’, ‘univers’	‘dark’, ‘galaxy’, ‘galaxies’, ‘plastic’, ‘black’, ‘light’, ‘universe’ ‘like’, ‘stars’, ‘space’	6	14	0.43
‘devons’, ‘monde’, ‘donc’, ‘cette’, ‘changement’, ‘climatique’, ‘missions’, ‘tout’, ‘probl’, ‘cela’	‘year’, ‘world’, ‘carbon’, ‘people’, ‘going’, ‘climate’, ‘global’, ‘need’, ‘change’, ‘energy’	5	16	0.31
‘trois’, ‘fois’, ‘moins’, ‘deux’, ‘comme’, ‘gens’, ‘monde’, ‘cette’, ‘jour’, ‘bien’	‘much’, ‘many’, ‘make’, ‘first’, ‘percent’, ‘system’, ‘people’, ‘years’, ‘world’, ‘water’	2	18	0.11
‘alors’, ‘comme’, ‘gens’, ‘autre’, ‘amour’, ‘cette’, ‘rires’, ‘quand’ ‘cela’, ‘tout’	‘want’, ‘life’, ‘know’, ‘really’, ‘like’, ‘going’, ‘laughter’, ‘people’, ‘think’, ‘love’	4	16	0.25
‘soleil’, ‘cela’, ‘donc’, ‘tout’, ‘mars’, ‘cette’, ‘surface’, ‘plan’ ‘comme’, ‘terre’	‘years’, ‘look’, ‘mars’, ‘life’, ‘earth’ ‘going’, ‘actually’, ‘like’, ‘planet’, ‘know’	4	16	0.25

Words which means the same in both languages are considered to be intersection between the two sets in our case. For example, ‘lumi’ which is the shortened word for ’lumiere’ in French, means ‘light’ in English which would be counted as intersection. There are some cases where a word might have multiple equivalents in the other language. For example, in topic 5 for DP-LBLA, the French word ‘monde’ can mean both ‘global and ‘world’ in English. In these cases we consider both the English words as intersection. Based on our analysis, we found that DP-LBLA had 24 similar words overall in the 5 topics averaging a Jaccard index of 0.33 whereas poly-LDA had only 21 words in common with an average Jaccard index of 0.27. In addition to these metrics, eye-balling the topics would clearly indicate the quality of topics derived by DP-LBLA.

Furthermore, since we found that using the interactive version, the quality of topic can be improved considerably form the Chapter 4, we apply the idea to improve the quality of these derived topics. The experiment was conducted

Table 6.5: Improvement in coherence score for DP-LBLA with interactive learning

Model (Language)	$\eta_1 = 0.2$	$\eta_1 = 0.4$	$\eta_1 = 0.6$	$\eta_1 = 0.8$	$\eta_1 = 1$
DP-LBLA (En)	-0.31	-0.32	-0.32	-0.35	-0.35
DP-LBLA (Fr)	-3.96	-3.96	-4.15	-4.39	-4.50

varying the weights for the objective and subjective probabilities. We can see that as we keep increasing the value of η_1, the coherence decreases. It is to be noted that, at $\eta_1 = 1$, the model acts as a regular DP-LBLA model. The observations clearly show the effect of varying the impact of user defined probabilities. The pattern is plainly visible in Fig. 6.5 and Fig. 6.6. In fig. 6.5 and Fig. 6.6, η denotes η_1 in general for simplicity. In case, the user modifying the probabilities is new to the topics involved in the documents, keeping a higher value for η_1 will help maintaining the performance of our model.

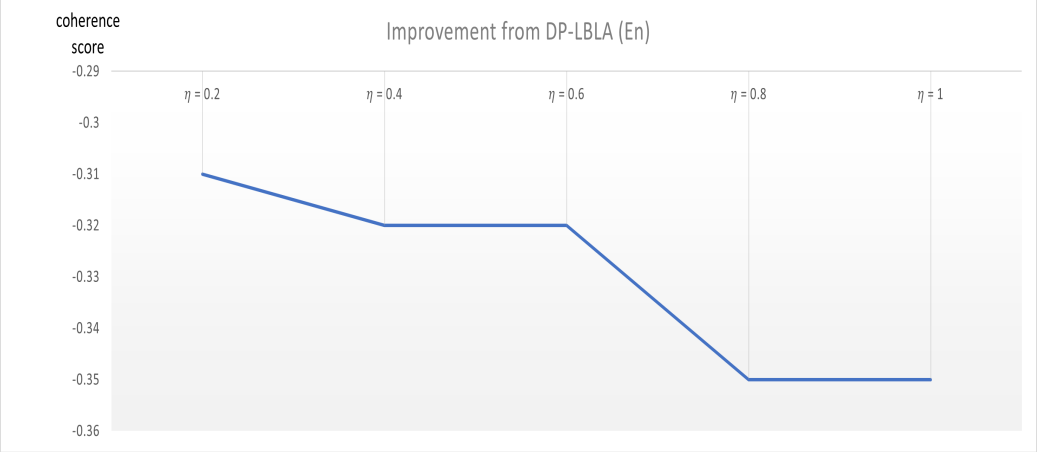

Figure 6.5: Improvement in coherence score from DP-LBLA with interactive learning for English topics

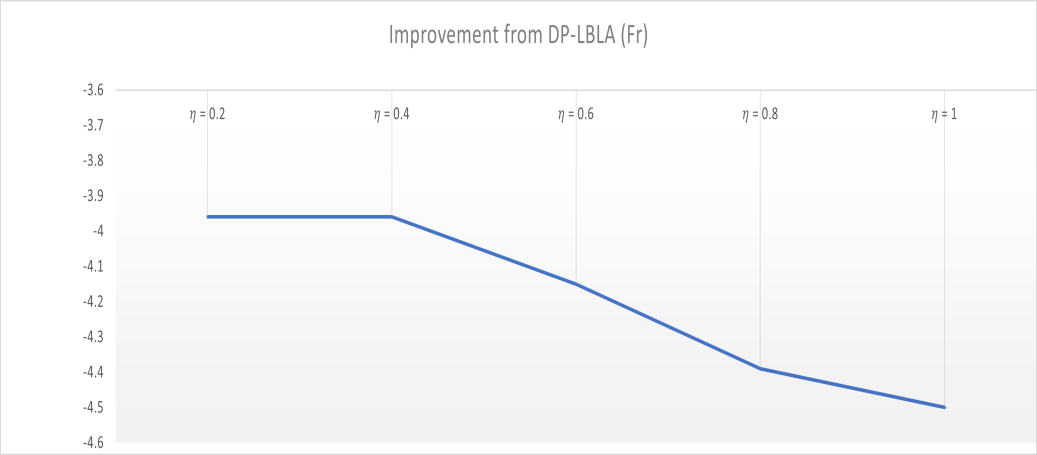

Figure 6.6: Improvement in coherence score from DP-LBLA with interactive learning for French topics

Chapter 7

Conclusion

In this thesis, we have proposed novel ideas to improve existing topic modelling approaches. We have proposed models that are flexible and can be used to tackle a number of practical tasks. A few of our applications show how the model can be easily tweaked to handle a specific problem. The adaptability of our models is quite evident from the experiments. A fine analysis has been carried out in each of the chapters to examine the effectiveness of our models with respect to varied applications.

In chapter 3, we have presented two novel topic models which can be used in a number of applications provided the input is count data. We have explained the mathematical model and provided a variational method to estimate the parameters both for batch processing and online use cases. We have also presented a way to convert the model into a supervised setting which learns the topics pertaining to classes simultaneously during training. The experiments were done with less documents for training to test how good the model performs in cases where data availability is low since this is the case in several industrial applications. The accuracy for various applications for classification stands proof for the efficiency of the models in detecting topics. Both models are found to perform really well and may provide efficient alternatives to the standard methods.

In Chapter 4, we have introduced an interactive algorithm, which can be used with our basic models to achieve better quality topics. The ability to tune the effect of user input is another useful asset that prevents us from wrongly modifying the topic probabilities. The experiments with two standard datasets prove that our model is capable of drawing quality topics

compared to the other baseline models with some input from the user. We also explored the flexibility of our model to control the impact of user inputs. The results indicate that our model could be very efficient in tasks involving unsupervised topic learning.

Chapters 5 and 6 show the flexibility of our model to be easily adopted for two interesting applications namely, recommendation systems and multilingual topic extraction. In the case of recommendation systems, from the example queries, we see that our models are able to deliver promising suggestions that the user might like. Using biterms in conjunction with our models tend to improve the results considerably. Especially, Bi-LBLMA model proves to be a good alternative to LDA based on the results from both experiments. Considering the results for English and French language documents in the case of multilingual topic extraction, both the models perform better than the baseline models LDA and poly-LDA. Though DP-LBLA performs a little better than DP-LGDA their metrics are still closely on par with each other. Our experiments also reveal the advantages of using GD and BL distributions in place of Dirichlet. The method also overcomes the drawback of manual component selection in our model. The performance boost achieved by interactive learning proves to be promising.

The overall performance of our model is found to be quite satisfactory as evident from the results. Future work may include improving the variational algorithm by relaxing the independency constraint by using a collapsed variational algorithm. A wide range of altered variational algorithms are available that could further speed up inference [104]. With the proliferation of ANN based models in multiple domains it would be an interesting approach to see how our models integrate with neural topic models (NTM). Using our models in conjunction with models like variational auto encoders (VAE) and generative adversarial networks (GANs) would be first steps to move forward in the scope of ANN.

Appendix A

Proof of Equations

The proof of equations for the variational solutions is explained in detail in this chapter. The procedure followed in these derivations can be extended to other solutions easily. According to Eq. 3.19, every term other than $Q_j(\Theta_j)$ is considered to be a constant. This variational solution can be found by taking the logarithmic form of Eq. 3.12 and Eq. 3.15. We illustrate the derivations for a few parameters in LGDMA model, which can be extended to the rest of the parameters and in a broader sense, to LBLMA as well.

Variational solution to $Q(z_{dnk})$

To find the variational solution for $Q(z_{dnk})$, let's gather the terms that contains z_{dnk} from Eq. 3.12. Taking the logarithm of these collected terms, we get,

$$\begin{aligned}\ln Q(z_{dnk}) &= \sum_{v=1}^{V} z_{dnk} w_{dnv} \ln \beta_{kv} + z_{dnk} \ln \theta_{dk} + const \\ &= \sum_{v=1}^{V} z_{dnk} \left[w_{dnv} \ln \beta_{kv} + \langle \ln \theta_{dk} \rangle \right] + const \end{aligned} \tag{A.1}$$

const in the equations indicate the rest of the parameters which are assumed to be constant due to independency. It is known that $w_{dnv} = 1$ only when the word in the vocabulary v is the same as respective word and hence we can rewrite the equation as,

$$\ln Q(z_{dnk}) = z_{dnk} \left[\ln \beta_{kv} + \langle \ln \theta_{dk} \rangle \right] + const \tag{A.2}$$

Let $ln\delta_{dnk} = \ln \beta_{kv} + \langle \ln \theta_{dk} \rangle$. This changes the equation to,

$$\ln Q(z_{dnk}) = z_{dnk} \ln \delta_{dnk} + const \tag{A.3}$$

which when exponentiated turns to,

$$Q(z_{dnk}) \propto \delta_{dnk}^{z_{dnk}} \tag{A.4}$$

Normalizing δ_{dnk} with $\phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K} \delta_{dnk}}$. We can write the final variational solution as,

$$Q(z_{dnk}) = \phi_{dnk}^{z_{dnk}} \tag{A.5}$$

since, this is the form of a multinomial distribution, we can write $\langle z_{dnk} \rangle = \phi_{dnk}$. Similarly, we can derive the equations corresponding to $Q(y_{dl})$ following the same method.

Variational solution for $Q(\vec{sigma})$

Similarly, taking the logarithmic terms involving σ_{lk}, we can write,

$$\ln Q(\sigma_{lk}) = \sum_{d=1}^{D} \langle y_{dl} \rangle \mathcal{F}_l + \sigma_{lk} \ln \theta_{dk} + (u_{lk} - 1) \ln \sigma_{lk} - \nu_{lk} \sigma_{lk} + const \tag{A.6}$$

provided, $\mathcal{F}_l = \left\langle \ln \frac{\Gamma(\sigma_{lk} + \tau_{lk})}{\Gamma(\sigma_{lk})\Gamma(\tau_{lk})} \right\rangle$. Since this expectation is not tractable we use an approximation similar to [105] which gives,

$$\mathcal{F} \geq \ln \sigma \left[(\psi(\overline{\sigma} + \overline{\tau}) - \psi(\overline{\sigma}) + \overline{\tau} \psi'(\overline{\sigma} + \overline{\tau})) \left(\langle \ln \tau \rangle - \ln \overline{\tau} \right) \right] \overline{\sigma} \tag{A.7}$$

substituting this in Eq. A.6 and collecting the similar terms, we can rewrite the equation for $\ln Q(\sigma_{lk})$ as,

$$\begin{aligned} \ln Q(\sigma_{lk}) = & \ln \sigma_{lk} \Bigg[\sum_{d=1}^{D} \langle y_{dl} \rangle \Big[\psi(\overline{\sigma}_{lk} + \overline{\tau}_{lk}) - \psi(\overline{\sigma}_{lk}) \\ & + \overline{\tau}_{lk} \psi'(\overline{\sigma}_{lk} + \overline{\tau}_{lk}) \Big] \left(\langle \ln \tau_{lk} \rangle - \ln \overline{\tau}_{lk} \right) \overline{\sigma}_{lk} + u_{jl} - 1 \Bigg] \\ & + \sigma_{lk} \left[\sum_{d=1}^{D} \langle y_{dl} \rangle \ln \theta_{dk} - \nu_{lk} \right] + const \end{aligned} \tag{A.8}$$

We can see that Eq. A.8 is the logarathmic form of Gamma distribution. Hence, exponentiating this equation will result in the variational solution,

$$Q(\sigma_{lk}) = \mathcal{G}(\sigma_{lk} \mid \upsilon^*_{lk}, \nu^*_{lk}) \tag{A.9}$$

where υ_{lk} and ν_{lk} is given by equations 3.29 and 3.31 respectively. Likewise, we can find the variational solutions for the rest of the parameters for LGDMA. The inference method for LBLMA also follows the same procedure.

List of References

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, Sep 01 1990. Last updated - 2013-02-24.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[3] Chunjun Zheng and Yi Zhang. Image retrieval based on lda and svm. In *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 446–449, 2019.

[4] Nikhil Rasiwasia and Nuno Vasconcelos. Latent dirichlet allocation models for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2665–2679, 2013.

[5] Zhibo Wang, Long Ma, and Yanqing Zhang. A hybrid document feature extraction method using latent dirichlet allocation and word2vec. In *2016 IEEE first international conference on data science in cyberspace (DSC)*, pages 98–103. IEEE, 2016.

[6] Xiaogang Wang and Eric Grimson. Spatial latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 20:1577, 2007.

[7] Hridoy Sankar Dutta, Vishal Raj Dutta, Aditya Adhikary, and Tanmoy Chakraborty. Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*, 15:2667–2678, 2020.

[8] John Lafferty and David Blei. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[9] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 2010.

[10] Jon Mcauliffe and David Blei. Supervised topic models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[11] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, 3:19–48, 2010.

[12] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[13] Nizar Bouguila and Walid ElGuebaly. On discrete data clustering. In Takashi Washio, Einoshin Suzuki, Kai Ming Ting, and Akihiro Inokuchi, editors, *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20-23, 2008 Proceedings*, volume 5012 of *Lecture Notes in Computer Science*, pages 503–510. Springer, 2008.

[14] Nizar Bouguila and Khalid Daoudi. Learning concepts from visual scenes using a binary probabilistic model. In *2009 IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, 2009.

[15] Reza Bahmanyar, Shiyong Cui, and Mihai Datcu. A comparative study of bag-of-words and bag-of-topics models of eo image patches. *IEEE Geoscience and Remote Sensing Letters*, 12(6):1357–1361, 2015.

[16] Yanshan Wang, Jae-Sung Lee, and In-Chan Choi. Indexing by latent dirichlet allocation and an ensemble model. *Journal of the Association for Information Science and Technology*, 67(7):1736–1750, 2016.

[17] Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. *IEEE*

Transactions on Knowledge and Data Engineering, 21(12):1649–1664, 2009.

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

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

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

[21] Ola Amayri and Nizar Bouguila. Content-based spam filtering using hybrid generative discriminative learning of both textual and visual features. In *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 862–865, 2012.

[22] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA, 1999. Association for Computing Machinery.

[23] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Novel mixtures based on the dirichlet distribution: Application to data and image classification. In Petra Perner and Azriel Rosenfeld, editors, *Machine Learning and Data Mining in Pattern Recognition*, pages 172–181, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[24] Nizar Bouguila, Djemel Ziou, and Riad I. Hammoud. A bayesian non-gaussian mixture analysis: Application to eye modeling. In *2007 IEEE*

Conference on Computer Vision and Pattern Recognition, pages 1–8, 2007.

[25] Nizar Bouguila and Wentao Fan. *Mixture models and applications*. Springer, 2020.

[26] Nuha Zamzami and Nizar Bouguila. Probabilistic modeling for frequency vectors using a flexible shifted-scaled dirichlet distribution prior. *ACM Trans. Knowl. Discov. Data*, 14(6):69:1–69:35, 2020.

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

[28] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

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

[30] Nizar Bouguila and Djemel Ziou. *A powerful finite mixture model based on the generalized Dirichlet distribution: unsupervised learning and applications*, volume Proceedings of the 17th International Conference on Pattern Recognition, ICPR, pages 280–283. 2004.

[31] Jen-Tzung Chien, Chao-Hsi Lee, and Zheng-Hua Tan. Latent dirichlet mixture model. *Neurocomputing*, 278:12–22, 2018.

[32] Wei Bian and Dacheng Tao. Dirichlet mixture allocation for multiclass document collections modeling. In *2009 Ninth IEEE International Conference on Data Mining*, pages 711–715. 9th IEEE International Conference on Data Mining, 2009.

[33] Carl Edward Rasmussen et al. The infinite gaussian mixture model. *Advances in Neural Information Processing Systems*, 12:554, 1999.

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

[35] Narges Manouchehri and Nizar Bouguila. Learning of finite two-dimensional beta mixture models. In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 227–232. 9th IEEE International Symposium on Signal, Image, Video and Communications (ISIVC), 2018.

[36] Jaspreet Singh Kalsi and Nizar Bouguila. Color image segmentation using generalized inverted dirichlet finite mixture models by integrating spatial information. In *28th IEEE International Symposium on Industrial Electronics, ISIE 2019, Vancouver, BC, Canada, June 12-14, 2019*, pages 1379–1384. IEEE, 2019.

[37] Xuanbo Su, Nizar Bouguila, and Nuha Zamzami. Covid-19 news clustering using mcmc-based learing of finite EMSD mixture models. In Eric Bell and Fazel Keshtkar, editors, *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*, 2021.

[38] Mohamad Mehdi, Nizar Bouguila, and Jamal Bentahar. Trustworthy web service selection using probabilistic models. In Carole A. Goble, Peter P. Chen, and Jia Zhang, editors, *2012 IEEE 19th International Conference on Web Services, Honolulu, HI, USA, June 24-29, 2012*, pages 17–24. IEEE Computer Society, 2012.

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

[40] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Expectation propagation learning of finite multivariate beta mixture models and applications. *Neural Comput. Appl.*, 34(17):14275–14285, 2022.

[41] Nizar Bouguila and Djemel Ziou. On fitting finite dirichlet mixture using ECM and MML. In Peng Wang, Maneesha Singh, Chidanand Apté, and Petra Perner, editors, *Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, 2005, Proceedings, Part I*, volume 3686 of *Lecture Notes in Computer Science*, pages 172–182. Springer, 2005.

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

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

[44] Yee Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[45] William M Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 642–647. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011.

[46] Chris Lloyd, Tom Gunter, Michael Osborne, Stephen Roberts, and Tom Nickson. Latent point process allocation. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 389–397, Cadiz, Spain, 09–11 May 2016. PMLR.

[47] Ali Shojaee Bakhtiari and Nizar Bouguila. A latent beta-liouville allocation model. *Expert Systems with Applications*, 45:260–272, 2016.

[48] Ali Shojaee Bakhtiari and Nizar Bouguila. Online learning for two novel latent topic models. In Linawati, Made Sudiana Mahendra, Erich J. Neuhold, A. Min Tjoa, and Ilsun You, editors, *Information and Communication Technology*, pages 286–295, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[49] Koffi Eddy Ihou and Nizar Bouguila. Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing*, 332:372–395, 2019.

[50] Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[51] Elise Epaillard and Nizar Bouguila. Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):1034–1047, 2019.

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

[53] N. Bouguila and D. Ziou. Mml-based approach for high-dimensional unsupervised learning using the generalized dirichlet mixture. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 53–53, 2005.

[54] Nizar Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, 2011.

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

[56] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1323–1329. IJCAI/AAAI, 2013.

[57] Yee Whye Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

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

[59] Koffi Eddy Ihou, Manar Amayri, and Nizar Bouguila. Stochastic variational optimization of a hierarchical dirichlet process latent beta-liouville topic model. *ACM Trans. Knowl. Discov. Data*, 16(5):84:1–84:48, 2022.

[60] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Proceedings of the Conference on Empirical Methods in Natural Language Processing; EMNLP '11, page 262–272, USA, 2011. Association for Computational Linguistics.

[61] Yezheng Liu, Fei Du, Jianshan Sun, and Yuanchun Jiang. ilda: An interactive latent dirichlet allocation model to improve topic quality. *Journal of Information Science*, 46(1):23–40, 2020.

[62] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017.

[63] Ali Shojaee Bakhtiari and Nizar Bouguila. A variational bayes model for count data learning and classification. *Engineering Applications of Artificial Intelligence*, 35:176–186, 2014.

[64] Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning for finite dirichlet mixture models and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 23(5):762–774, 2012.

[65] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, Cambridge, Massachusetts, 2001.

[66] Nizar Bouguila. Deriving kernels from generalized dirichlet mixture models and applications. *Information Processing and Management*, 49(1):123–137, 2013.

[67] Masaaki Sato. Online Model Selection Based on the Variational Bayes. *Neural Computation*, 13(7):1649–1681, 07 2001.

[68] Zelong Liu, Maozhen Li, Yang Liu, and Mahesh Ponraj. Performance evaluation of latent dirichlet allocation in text mining. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 2695–2698. IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[69] Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910, 2009.

[70] Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo, and Alfonso Urso. Genomic sequence classification using probabilistic topic modeling. In Enrico Formenti, Roberto Tagliaferri, and Ernst Wit, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 49–61, Cham, 2014. Springer International Publishing.

[71] Michael R. McLaren and Benjamin J.) Callahan. Silva 138.1 prokaryotic ssu taxonomic training data formatted for dada2 [data set] (2021) zenodo.

[72] Paul E Rybski, Daniel Huber, Daniel D Morris, and Regis Hoffman. Visual classification of coarse vehicle orientation using histogram of oriented gradients features. In *2010 IEEE Intelligent vehicles symposium*, pages 921–928. IEEE, 2010.

[73] Yuyao Wang, Zhengming Li, Long Wang, Min Wang, et al. A scale invariant feature transform based method. *J. Inf. Hiding Multim. Signal Process.*, 4(2):73–89, 2013.

[74] Guang-Hai Liu, Jing-Yu Yang, and ZuoYong Li. Content-based image retrieval using computational visual attention model. *Pattern Recognition*, 48(8):2554–2566, 2015.

[75] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[76] Jian Yang, Qi Zhang, Xiaofeng Jiang, Shuangwu Chen, and Feng Yang. Poirot: Causal correlation aided semantic analysis for advanced persistent threat detection. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[77] Ziqiang Wang and Xu Qian. Text categorization based on lda and svm. In *2008 International Conference on Computer Science and Software Engineering*, volume 1, pages 674–677. IEEE, 2008.

[78] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[79] Michael J. Pazzani and Daniel Billsus. *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[80] Jesus Bobadilla, Antonio Hernando, Fernando Ortega, and Jesus Bernal. A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38(12):14609–14623, 2011.

[81] Rohit Nagori and G. Aghila. Lda based integrated document recommendation model for e-learning systems. In *2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pages 230–233, 2011.

[82] Dhiraj Vaibhav Bagul and Sunita Barve. A novel content-based recommendation approach based on lda topic modeling for literature recommendation. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 954–961, 2021.

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

[84] Nuha Zamzami and Nizar Bouguila. Mml-based approach for determining the number of topics in EDCM mixture models. In Ebrahim Bagheri and Jackie Chi Kit Cheung, editors, *Advances in Artificial Intelligence - 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8-11, 2018, Proceedings*, volume 10832 of *Lecture Notes in Computer Science*, pages 211–217. Springer, 2018.

[85] E Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60, 2016.

[86] Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 14(3):1–22, 2015.

[87] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147, 2015.

[88] Jagadeesh Jagarlamudi and Hal Daumé. Extracting multilingual topics from unaligned comparable corpora. In *European Conference on Information Retrieval*, pages 444–456. European Conference on Information Retrieval, 2010.

[89] Shudong Hao and Michael Paul. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th international conference on computational linguistics*, pages 2595–2609. Association for Computational Linguistics, 2018.

[90] Michelle Yuan, Benjamin Van Durme, and Jordan L. Ying. Multilingual anchoring: Interactive topic modeling and alignment across languages. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8667–8677, 2018.

[91] Ueli Reber. Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures*, 13(2):102–125, 2019.

[92] Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1243–1248. Association for Computational Linguistics, 2019.

[93] Elaine Zosa and Mark Granroth-Wilding. Multilingual dynamic topic model. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, 2019*, pages 1388–1396, Varna, Bulgaria, 2019. INCOMA Ltd.

[94] Alexander Böhm, Stefan Reiners-Selbach, Jan Baedke, Alejandro Fábregas Tejeda, and Daniel J. Nicholson. What was theoretical biology? A topic-modelling analysis of a multilingual corpus of monographs and journals, 1914-1945. In Michaela Geierhos, editor, *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2022, Potsdam, Germany, March 7 - 11, 2022*, 2022.

[95] Elaine Zosa, Lidia Pivovarova, Michele Boggia, and Sardana Ivanova. Multilingual topic labelling of news topics using ontological mapping. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 248–256, Varna, Bulgaria. Springer.

[96] Bagher Rahimpour Cami, Hamid Hassanpour, and Hoda Mashayekhi. User preferences modeling using dirichlet process mixture model for a content-based recommender system. *Knowledge-Based Systems*, 163:644–655, 2019.

[97] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):256–270, 2014.

[98] Carl E Rasmussen and Zoubin Ghahramani. *Infinite mixtures of Gaussian process*, page Advances in Neural Information Processing Systems. 2002.

[99] Hirotugu Akaike. Factor analysis and aic. *Psychometrika*, 52(3):317–332, 1987.

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

[101] Jen-Tzung Chien, Chao-Hsi Lee, and Zheng-Hua Tan. *Dirichlet mixture allocation*, page IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). 2016.

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

[103] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, page 880–889, USA, 2009. Association for Computational Linguistics.

[104] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.

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