Advancements in Topic Modeling: Integrating Bi-Directional Recurrent Attentional Models, Neural Embeddings, and Flexible Distributions

Pantea Koochemeshkian

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Information and Systems Engineering) at

Concordia University

Montréal, Québec, Canada

September 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Pantea Koochemeshkian**

Entitled: **Advancements in Topic Modeling: Integrating Bi-Directional Recurrent Attentional Models, Neural Embeddings, and Flexible Distributions**

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

Doctor of Philosophy (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:

_____ Chair
Dr. Sang Hyeok Han

_____ External Examiner
Dr. Thar Baker Shamsa

_____ External to Program
Dr. Mazdak Nik-Bakht

_____ Examiner
Dr. Chun Wang

_____ Examiner
Dr. Suryadipta Majumdar

_____ Supervisor
Dr. Nizar Bouguila

Approved by _____
Dr. Chun Wang, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

_____ 2024

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

Abstract

Advancements in Topic Modeling: Integrating Bi-Directional Recurrent Attentional Models, Neural Embeddings, and Flexible Distributions

Pantea Koochemeshkian, Ph.D.

Concordia University, 2024

A primary objective in natural language processing is the classification of texts into discrete categories. Topic models and mixture models are indispensable tools for this task, as they both acquire patterns from data in an unsupervised manner. Several extensions to established topic modeling frameworks are introduced by incorporating more flexible priors and advanced inference methods to enhance performance in text document analysis. The Multinomial Principal Component Analysis (MPCA) framework, a Dirichlet-based model, is extended by integrating generalized Dirichlet (GD) and Beta-Liouville (BL) distributions, resulting in GDMPCA and BLMPCA models. These priors address the limitations of the Dirichlet prior, such as its independent assumption within components and restricted covariance structure. Efficiency is further improved by implementing variational Bayesian inference and collapsed Gibbs sampling for fast and accurate parameter estimation.

Enhancements to the Bi-Directional Recurrent Attentional Topic Model (bi-RATM) are made by incorporating GD and BL distributions, leading to GD-bi-RATM and BL-bi-RATM models. These models leverage attention mechanisms to model relationships between sentences, offering higher flexibility and improved performance in document embedding tasks.

Extensions to the Dirichlet Multinomial Regression (DMR) and deep Dirichlet Multinomial Regression (dDMR) approaches are achieved by incorporating GD and BL distributions. This integration addresses limitations related to handling complex data structures and overfitting, with collapsed Gibbs sampling providing an efficient method for parameter inference. Experimental results on benchmark datasets demonstrate enhanced topic modeling performance, particularly in handling

complex data structures and reducing overfitting.

Novel approaches are developed by integrating embeddings derived from Bert-Topic with the multi-grain clustering topic model (MGCTM). Recognizing the hierarchical and multi-scale nature of topics, these methods utilize MGCTM to capture topic structures at multiple levels of granularity. By incorporating GD and BL distributions, the expressiveness and flexibility of MGCTM are enhanced. Experiments on various datasets show superior topic coherence and granularity compared to state-of-the-art methods.

Overall, the proposed models exhibit improved interpretability and effectiveness in various natural language processing and machine learning applications, showcasing the potential of combining neural embeddings with advanced probabilistic modeling techniques.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Nizar Bouguila, for his invaluable guidance, support, and encouragement throughout my PhD journey, he has always led and motivated me with endless patience. I am deeply grateful for his unwavering guidance and support. Meeting him as the instructor of the Data Mining course at Concordia University was a turning point in my academic journey. His expertise and insights have been instrumental in shaping this dissertation, and I am profoundly grateful for his mentorship.

I am also deeply thankful to my parents for their unwavering love, patience, and support. Their encouragement has been a constant source of strength and motivation, and I could not have completed this work without their belief in me.

Additionally, I would like to extend my heartfelt thanks to my lab mates: Hussain Albazza, Kamal, Eddy, Omar, Basim, Fatma, whose collaboration and camaraderie have made this journey enjoyable and intellectually stimulating. Their feedback and discussions have enriched my research experience, and I am grateful for their friendship and support.

Contents

List of Figures

List of Tables

List of Acronyms

Abbreviation	Full Term
BL	Beta-Liouville
BLMPCA	Beta-Liouille Multinomial Principal Component Analysis
BoW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
BioNLP	Biological Natural Language Processing
CGS	Collapsed Gibbs Sampling
DMR	Dirichlet Multinomial Regression
dDMR	Deep Dirichlet Multinomial Regression
DP	Dirichlet Process
GD	Generalized Dirichlet
GDMPCA	Generalized Dirichlet Multinomial Principal Component Analysis
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA	Latent Dirichlet Allocation
LSTM	Long Short-Term Memory
MCMC	Markov Chain Monte Carlo
MGCTM	Multi-Grain Clustering Topic Model
MPCA	Multinomial Principal Component Analysis
NMF	Non-negative Matrix Factorization
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic Analysis
pLSI	Probabilistic Latent Semantic Indexing
RNN	Recurrent Neural Network
SVD	Singular Value Decomposition
SVM	Support Vector Machine
UMAP	Uniform Manifold Approximation and Projection
dGDMR	Deep Generalized Dirichlet Multinomial Regression
dBLMR	Deep Beta-Liouville Multinomial Regression
BLMR	Beta-Liouville Multinomial Regression
EM	Expectation-Maximization

Chapter 1

Introduction

1.1 Introduction

In the fast-paced world of technological advances, the emergence of various digital data forms has significantly opened numerous opportunities for gathering valuable information. Every day, massive amounts of digital data, including a substantial portion of textual data, are stored in archives and available on the Internet, necessitating the development of effective and scalable statistical models to extract hidden knowledge from such rich data sources [1]. Advances in information technology combined with social media, where people now share knowledge and skills, have created an information revolution. Daily created websites are becoming great resources for data science and analytics, in addition to being stores of important information generally hidden in documents. One crucial task in machine learning is modeling documents into a vector space, making it essential to carefully analyze document components, including words, phrases, and paragraphs. Due to the complexity and variability of massive data collections, processing such unstructured records necessitates efficient machine learning techniques. In topic modeling, such collections are summarized as documents that use the bag-of-words method [2] to perform on count data, aiming to efficiently generate topics to make accurate predictions on unseen documents in tasks like retrieval and classification.

One of the main challenges in the statistical analysis of textual data is capturing and representing its complexity. Various approaches have been applied to address this issue, leveraging the rapid

development of information technology. Vast quantities of scientific documents are now freely available for mining, making the analysis and mining of these documents active research areas for many years. Data projection and clustering are crucial for document analysis. Projection aims at creating low-dimensional, meaningful data representations, while clustering groups similar data patterns [3, 4]. Traditionally, these methods have been studied separately, but they intersect in many applications [4]. K-means clustering, although widely used for creating compact cluster representations, does not fully capture document semantics. This gap has led to the adoption of machine learning and deep learning techniques to address text mining challenges, including text classification [5], summarization [6], segmentation [7], topic modeling [8], and sentiment analysis [9].

In this thesis, we focus specifically on topic modeling aspects. Topic models are generally classified into two categories: those based on matrix decomposition, like singular value decomposition (SVD), and generative models [10]. The matrix decomposition approach, such as probabilistic latent semantic analysis (PLSA) [11, 12], analyzes text by mining and requires a deep understanding of the corpora structure. PLSA, also known as probabilistic latent semantic indexing (pLSI) [12], represents documents as a mix of topics by performing matrix decomposition on the term-document matrix and is effective in identifying relevant words for each topic. In contrast, the generative approach of topic modeling focuses on the context of words across the entire document corpus. These models use latent variable models, treating a document as a combination of various topics, each represented by a random vector of words [4].

However, research by [13] indicates that while the probabilistic latent semantic indexing (pLSI) model offers some insights, it falls short in clustering and as a generative model due to its inability to generalize to new documents. To address these limitations, Latent Dirichlet Allocation (LDA) [13] was introduced, enhancing pLSI by using Dirichlet distribution for topic mixtures. LDA stands out as a more effective generative model, though it still lacks robust clustering capabilities [4]. The integration of clustering and projection into a single framework has been a recent focus in this field, recognizing the need to combine these two approaches [14, 15].

The main issue with current text analysis models is their failure to clearly define a probability model encompassing hidden variables and assumptions [12, 16, 17, 18]. To address this, variational Expectation-Maximization (EM) has been utilized, notably in Multinomial PCA (MPCA),

which links topics to latent mixture proportions in a probabilistic matrix factorization framework [18, 19]. Extensions of LDA, like its hierarchical [20] and online versions [21], have been developed, although they lack the integration of Dirichlet priors in modeling. Researchers have explored alternative models using conjugate priors and methods like Gibbs Sampling and Markov Chain Monte Carlo (MCMC) [22], which, despite their effectiveness, require longer convergence times compared to the variational Bayes approach.

A significant constraint in topic modeling is the reliance on the bag-of-words framework, which simplifies texts to mere word counts, often sidelining the nuanced semantic connections that exist between words. This limitation has led to the adoption of text embedding techniques such as Bidirectional Encoder Representations from Transformers (BERT), which generates contextually enriched vector representations of words and sentences, capturing semantic intricacies and allowing for deeper topic discernment [23, 24].

Embedding techniques have permeated various natural language processing tasks, from classification to powering sophisticated neural search engines. Their application in topic modeling has drawn increasing attention, with methods like Top2Vec, which uses document clustering to identify central themes, demonstrating the effectiveness of embedding techniques in representing topics [25, 26]. The integration of embeddings with topic models has shown significant improvements in capturing the nuanced semantics of text data.

The intricate domains of document clustering and topic modeling share a symbiotic relationship, each enriching the other's capabilities. Topic models discern subtle, underlying semantics in documents, offering a refined lens for delineating distinct document groups. This process involves transforming the document corpus into a topic-centric space, effectively filtering out extraneous noise and allowing for more precise and coherent clustering [4]. Conversely, clustering documents can reveal both local and global topics, enhancing the overall understanding of the document collection [27, 28].

Despite these advancements, traditional topic modeling methods often fail to capture the intricate dynamics within document groups. Current models lack the capacity to encapsulate evolving dynamics within document collections, yielding generalized topics that obscure both local and

global thematic nuances [13, 12, 16]. A more sophisticated approach acknowledges the symbiotic relationship between document clustering and topic modeling, leveraging enhanced clustering methodologies to inform more precise topic modeling outcomes [29].

1.2 Motivation for topic Modeling

Assume that our goal is to identify the main patterns found in the given text sample. A highly efficient and versatile approach to text analysis involves examining the frequencies of words within the text [30]. This method, known as frequency analysis, counts the number of times each word appears in the document, providing a foundational understanding of the text's content and structure.

Consider the following paragraph:

Chapter 2

Background and Preliminary Concepts

2.1 Related Work

In this section we explore the extensive range of literature on different approaches to topic modeling. The basis of this discipline relies on conventional topic modeling techniques [11, 12], but has also been greatly influenced by topic-class modeling [31, 32, 33] and the detailed examination of global and local document characteristics [34, 35].

Innovative strides have been made with the introduction of a two-stage topic extraction model for bibliometric data analysis, employing word embeddings and clustering for a more refined topic analysis [36]. This approach provides a nuanced lens to view the thematic undercurrents of scholarly communication. In recent times, numerous models have emerged to handle diverse types of data such as videos, images, and documents in various tasks, such as object detection, content detection, data management, and representation learning for words, phrases, and texts [37]. Undirected graphical models and directed graphical models are the most used representation learning approaches for documents. The replicated softmax model (RSM) produces distributed representations of texts using a two-layer undirected graphical model [38]. Many neural network-based techniques have recently been presented to handle collections of unlabeled documents or data containing different modalities [39, 37].

The landscape of sentiment analysis is similarly evolving, with breakthroughs like a term-weighted neural language model paired with a stacked bidirectional LSTM (Long short-term memory) framework, enhancing the detection of subtle sentiments like sarcasm in text [40]. Such advancements offer deeper insights into the complexities of language and its sentiments.

Cross-modal sentiment analysis also takes center stage with deep learning techniques, as seen in works that identify emotions from facial expressions [41, 42]. These studies, which utilize convolutional neural networks and Inception-V3 transfer learning [43], pave the way for multimodal sentiment analysis, potentially influencing strategies for textual sentiment analysis.

A hybrid deep learning method has been introduced to analyze sentiment polarities and knowledge graph representations, particularly focusing on health-related social media data, like tweets on monkeypox [44]. This underscores the importance of versatile and dynamic models in interpreting sentiment from real-time data streams.

Collectively, these contemporary works highlight the expansive applicability and dynamic nature of deep learning across various domains and data types. Their inclusion in our review underlines the potential for future cross-disciplinary research, expanding the scope of sentiment analysis to include both text and image data.

Alongside these emerging approaches, well-established techniques such as principal component analysis (PCA) and its text retrieval counterpart, latent semantic indexing [45], continue to be pivotal. Probabilistic latent semantic indexing (pLSI) [12] and Latent Dirichlet allocation (LDA) [13] further enrich the discussion on discrete data and topic modeling. Non-negative matrix factorization (NMF) [16] has also demonstrated effectiveness, emphasizing the need for models that can simultaneously handle clustering and projection. Addressing a gap in the literature, a multinomial PCA model has been proposed to offer probabilistic interpretations of the relationships between documents, clusters, and factors [18].

Moreover, several probabilistic topic models have been proposed as directed graphical models, and they have been employed to address various challenges. Some of these models include Latent Dirichlet Allocation (LDA) [13], Latent Generalized Dirichlet Allocation (LGDA) [46, 47], Latent Beta-Liouville Allocation (LBA) [48], and Correlated Topic Model (CTM) [49]. They are used to represent unstructured documents under the assumption that the words in a document arise from a

mixture of latent topics, with each topic being a distribution over the vocabulary [37].

Existing topic models only address word-level document sequentially. Sequential LDA [50] evaluates a document's sequential structure using a hierarchical two-parameter Poisson–Dirichlet process [51].

Many academics are interested in the lifelong topic model [52, 53, 54]. RNNs [55] are an effective method for dealing with sequentiality for texts at the word and sentence levels [56, 57]. The recurrent mechanism implemented in neural networks has achieved remarkable success in the field of language modeling. However, few subject modeling efforts have investigated the recurring mechanism. Several Bayesian models [58, 59] employ the recurrent process to address topic dynamics.

Much research [60, 61] integrates Bayesian models with deep neural networks. These models are considered deep Bayesian models, and they work with text directly, without considering more comprehensive details of the text, such as the word sequence in a document. Other studies use neural networks to generate topic models [62, 63], although these models require word embeddings.

Though few Bayesian models concentrate on language modeling with attention signals, all of the models that utilize attention mechanisms are neural network-based. Therefore, the authors of [37] introduced the attention mechanism into the bi-directional recurrent Bayesian topic model for documents.

This innovation is significant as it merges the strengths of traditional probabilistic models with modern neural network-based attention mechanisms, potentially enhancing the interpretability and performance of topic modeling. Directed graphical models have been widely used as probabilistic topic models to solve various problems. Latent Dirichlet allocation (LDA) [13], correlated topic model (CTM) [49], collapsed Gibbs sampler scheme for latent GD allocation (CGS-LGDA) [64], and collapsed Gibbs sampling Beta-Liouville multinomial (CGSBLM) [65] are examples of such models. These models and their adaptations are employed to model unstructured documents by assuming that the words in the document are a mixture of latent topics, and each topic is a distribution over the vocabulary [37].

The authors in [66] proposed sLDA, a supervised topic model that integrates topic and log-linear models. Instead of conditioning observed features and predicting topic variables, sLDA uses topic variables as inputs to the log-linear model to create observed features. sLDA needs to estimate

probability distributions over all possible feature values by fully specifying the link and dispersion functions for a GLM, which adds modeling complexity. To tackle the issue, Mimno and Andrew [67] proposed the Drichlet multinomial regression (DMR) topic model. DMR has an advantage over sLDA in many applications because it is fully conditional with respect to the observed features. DMR topic models are readily available and can be used with any set of features without requiring additional model specifications. Moreover, training a DMR model with complex and dependent features is as simple as training a model with a single real-valued feature.

As neural networks have become widely used, researchers have explored integrating topic models with neural models. One approach involves replacing traditional generative, LDA-based topic models with discriminatively trained models based on neural networks [68]. For instance, the authors in [68] and [69] use neural networks with softmax output layers and learn network parameters to maximize data likelihood. The authors of [69] also learn n-gram embeddings to identify topics whose elements are not limited to unigrams. Wan [70] takes a similar approach to dDMR, using a neural network to extract image representations for image classification, not document exploration as is typical of topic models. These models avoid approximating the posterior distribution of topic assignments given tokens by dropping the assumption that topic and word distributions are drawn from Dirichlet priors. In contrast, this study employs neural networks to learn feature representations for documents while keeping the core of the topic model unchanged, making dDMR agnostic to many other LDA extensions [71]. The integration of neural networks with topic models represents a significant advancement in the field, as it allows for the capture of more complex patterns in data. However, this approach still maintains the essence of traditional topic models, ensuring that the foundational principles of topic discovery remain intact. This blend of old and new techniques exemplifies the ongoing evolution in the field of natural language processing, where combining methods can yield superior results. Topic modeling has been a fundamental tool in natural language processing, enabling the discovery of latent themes within large collections of text. Traditional models such as Latent Dirichlet Allocation (LDA) [13] and NMF [72] describe documents as mixtures of latent topics, with each topic characterized by a distribution over words. However, these models often rely on bag-of-words representations, which fail to capture the semantic relationships between words, potentially limiting their ability to represent documents accurately [73].

To address the limitations of conventional models, recent advancements have incorporated neural embeddings to enhance topic modeling. Embedding techniques, such as Bidirectional Encoder Representations from Transformers (BERT) [23, 24], generate contextualized word and sentence embeddings, capturing semantic nuances more effectively than bag-of-words approaches. Models like Top2Vec and those proposed by [25] leverage these embeddings to cluster documents and extract topics by identifying words that are semantically close to cluster centroids.

BERTopic represents a significant advancement in this field by integrating state-of-the-art embedding techniques with topic modeling [24]. The model uses pre-trained transformer-based language models to generate document embeddings, clusters these embeddings, and extracts coherent topic representations through a class-based TF-IDF procedure [24]. BERTopic's workflow involves three key steps: generating document embeddings with Sentence-BERT (SBERT) [74], reducing the dimensionality of these embeddings using Uniform Manifold Approximation and Projection (UMAP) [75], and clustering the reduced embeddings with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [76, 77]. This approach allows for the creation of flexible and dynamic topic models that can be adjusted based on the embedding techniques and clustering methods used.

The synergy between document clustering and topic modeling has been explored in models such as MGCTM [27]. MGCTM integrates document clustering with topic modeling, discovering both global topics shared across clusters and local topics specific to each cluster. This integration allows for the simultaneous enhancement of clustering and topic modeling, as the discovery of latent groups in the document collection facilitates more accurate topic extraction and vice versa.

Compared to traditional models, both BERTopic and MGCTM demonstrate significant improvements in generating coherent and semantically meaningful topics. By leveraging modern embedding techniques and integrating clustering with topic modeling, these models enhance topic coherence and provide more representative topics aligned with the underlying document clusters. This integration is crucial for applications requiring accurate topic discovery and document organization, such as document summarization, classification, and retrieval [24, 27].

The evolution of topic modeling techniques, as exemplified by BERTopic and MGCTM, underscores the importance of integrating advanced neural embeddings and clustering methods to improve

the semantic coherence and representational accuracy of topics. These advancements highlight the potential for further research to enhance the interplay between clustering and topic modeling, leveraging the latest developments in machine learning and natural language processing.

2.2 Preliminary Concepts

2.2.1 Multinomial PCA

Our focus on the MPCA model and its extensions aims to consolidate these disparate strands of research, presenting a comprehensive framework for topic modeling that accounts for both clustering and projection, and reflecting the ongoing dialogue within the research community on these topics.

Multinomial PCA Topic model

Probabilistic approaches to reducing dimensions generally hypothesize that each observation x_i corresponds to a hidden variable, referred to as a latent variable θ_i. This latent variable exists within a subspace of dimension K. Typically, the relationship involves a linear mapping (β) within the latent space coupled with a probabilistic mechanism.

In the probabilistic PCA (pPCA) framework, as detailed in the work by [78], it is posited that each observation x_i originates from a standard Gaussian distribution $N_K(0_K; Z_K)$. The assumption of a Gaussian distribution is also employed for the conditional distribution of the observations.

$$x_i|\theta_i \sim \mathcal{N}_v(\beta\theta_i + \mu, \sigma^2 \mathcal{Z}_V) \tag{1}$$

where Z is a "standard" normal distribution, (β, μ) are the model parameters, and σ^2 is the variance that are learned by maximum likelihood inference.

The Gaussian assumption is suitable for real-valued data, yet it is less applicable to non-negative count data. Addressing this, [18] introduced a variant of pPCA where the latent variables are modeled as a discrete probability distribution, specifically using a Dirichlet distribution, where as $m \sim Dir(\alpha)$:

$$\mathcal{D}(m;\alpha) = \frac{1}{Z(\alpha)} \prod_{k=1}^{K} m_k^{\alpha_k - 1}(m) \tag{2}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k) \geq 0$.

Next, the probabilistic function is presumed to follow a multinomial distribution:

- m is modeled as a Dirichlet distribution with parameter α, representing a vector of probabilities that sum to one.
- C is drawn from a Multinomial distribution with m as the probability vector and L as the number of trials, indicating counts of various outcomes.
- Each w_k follows a Multinomial distribution with Ω_k as the probability vector and c_k the number of trials, counting the outcomes based on Ω_k.

The variables m and w assumed as hidden parameters for each document. For the parameter estimation of MPCA, first the variable Ω was estimated by the Dirichlet prior on m using parameters α [18]. The likelihood model for the MPCA is given as follows [19]:

$$p(m, w|\alpha, \Omega) = \frac{\eta(\sum_k \alpha_k)}{\prod_k \eta(\alpha_k)} C_{w_{1,1}, w_{1,2} \ldots} \prod_K m_K^{a_k - 1} \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \tag{3}$$

In the MPCA model, each observation x_i is conceptualized as a probabilistic combination of K topics that encapsulate the entire corpus. Here, m represents the mixture weights for the observation within the latent space, and Ω acts as a global parameter that contains all the corpus-level information.

As a result, the following equation is derived when the hidden variables have Dirichlet prior [18]:

$$\begin{aligned} m &\sim Dirichlet(\alpha) \\ \Omega_k &\sim Dirichlet(2f) \end{aligned} \tag{4}$$

The following updated formula converges to the local maximum, where $\frac{\eta(\sum_k \alpha_k)}{\prod_k \eta(\alpha_k)}$ is a normalizing

constant for the Dirichlet and r is the total row wise number of words in the document representation with the k component [18]:

$$\eta_{j,k,[i]} = \frac{\eta(\sum_k \alpha_k)}{\prod_k \eta(\alpha_k)} \frac{1}{\Omega_{k,j} m_{k,[i]}} \tag{5}$$

$$m_{k,[i]} = \frac{\eta(\sum_k \alpha_k)}{\prod_k \eta(\alpha_k)} \big(-1 + \alpha_k + \sum_j \eta_{j,k,[i]} r_{j,[i]} \big) \tag{6}$$

Eqs. 7 and 8 are the parameters for a multinomial and a Dirichlet respectively.

$$\Omega_{k,j} = \frac{\eta(\sum_k \alpha_k)}{\prod_k \eta(\alpha_k)} \big(2f + \sum_i \eta_{j,k,[i]} r_{j,[i]} \big) \tag{7}$$

$$\Psi(a_k) - \Psi(\sum_k a_k) = \frac{\log(1/k) + \sum_i \log(m_{k,[i]})}{1 + I} \tag{8}$$

According to exponential family definition (Appendix 8.1), Eq. 8 reformulates α using its dual representation. Minka's approach is used to derive α,where n_k is the number of times the outcome was k [79]:

$$\begin{aligned} n_k &= \sum_i \vartheta(x_i - k) \\ n_i &= \sum_k n_i k \end{aligned} \tag{9}$$

$$\alpha_k^{new} = a_k \frac{\sum_i \Psi(n_{ik} + a_k) - \Psi(a_k)}{\sum_i \Psi(n_i k + \sum_k a_k) - \Psi(\sum_k a_k)} \tag{10}$$

Connection between MPCA and LDA

The multinomial PCA model is closely connected to LDA [13] and forms the foundation over several topic models.

In text analysis, an observation typically refers to a document represented by a sequence of tokens or words, denoted as $w_i = w_{in}, n = 1 \dots L_i$. Each word w_{in} within a document i is initially

linked to a topic, which is specified by a vector z_{in} that is derived from a $Multinomial(1, \beta_k)$ distribution. The model for any given document i can be described as follows:

$$\begin{aligned} \theta_k &\sim Dirichlet(\alpha) \\ z|\theta &\sim Multinomial(1, \theta) \\ w|z &\sim Multinomial(1, \beta_k) \end{aligned} \tag{11}$$

At the word-level, marginalizing on z_{in} yields a distribution :

$$w_{in}|\theta_i \sim Multinomial(1, \beta\theta_i) \tag{12}$$

Furthermore, the distinction between LDA and MPCA is that LDA is a word-level model, whereas MPCA is a document-level model. Since the GDMPCA and BLMPCA are new variations of the MPCA, both new models are assumed to be document-level in the following proposed approaches.

2.2.2 Bi-Directional Recurrent Attentional Topic Model

Topic models usually assume that words are interchangeable, which is beneficial for efficient inference on large datasets [80]. Consequently, certain research has represented text as a sequence of words, using techniques such as n-gram language modeling [81] and recurrent neural networks for language modeling [82, 55, 37].

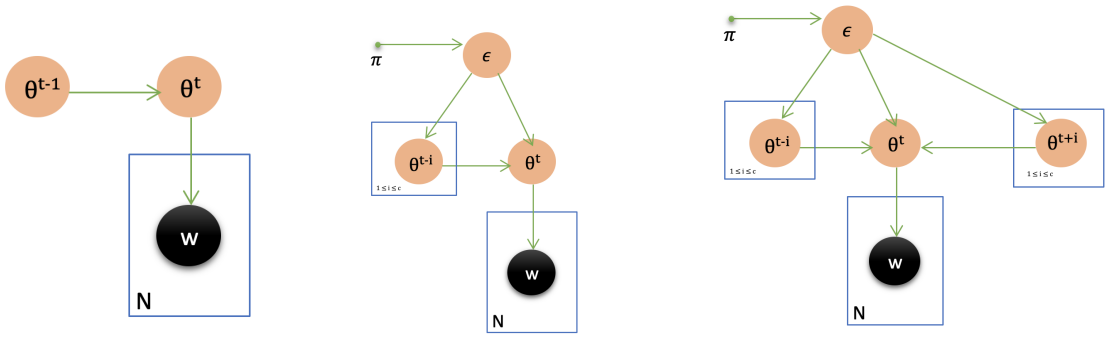
The bi-directional Recurrent Attention Topic Model (bi-RATM) is capable of modeling sequences of sentences using attention signals and taking into account sentence dependencies from two directions. To extend the model's ability to handle sequential data, a new model called bi-directional Recurrent Attention Bayesian Process (bi-RABP) has been proposed. The bi-RABP combines the Bayesian process with an attention mechanism to describe sequential data using recurrent local information. This model is particularly useful for analyzing bi-directional text data sequences. [37] also proposed a bi-RATM model to represent a probabilistic language model of bi-directional sentence sequences. The bi-RATM model, which employs bi-directional phrases, was introduced to acquire superior quality document representations. The performance of the proposed

model was evaluated through experiments, which demonstrated its effectiveness in document modeling and classification tasks. The bi-RATM model seamlessly integrates recurrent Bayesian techniques with attention signals, thereby enabling the adaptive learning of two-directional sentence sequences. The Bayesian attention technique can improve the modeling capacity of documents with bidirectional sentence sequences [37]. To handle large documents, efficient variational inference and an online bi-RATM algorithm have been devised. This approach allows bi-RATM to handle stream documents, making it useful for large corpora [37].

2.2.3 bi-RABP Model

Recurrent Chinese Restaurant Process

The Recurrent Chinese Restaurant Process (RCRP) [58], an extension of the Dirichlet Process, is a distribution over a Dirichlet distribution that models temporal coherence and variation of distributions over time in documents [58, 83]. In clustering models, the RCRP can be used as a prior for parameters, including a combination of the RCRP and the bag-of-words model for document modeling, as illustrated in Fig 2.1a. However, the prior's importance at the current time needs to be addressed in this approach. To tackle this issue, the Recurrent Attentional Bayesian Process (RABP) [37] is presented.



(a) RCRP Model. (b) RABP Model. (c) bi-RABP Model.

Figure 2.1: RCRP, RABP and bi-RABP models with the bag-of-words assumption.

Recurrent Attentional Bayesian Process

In [37], the Recurrent Attentional Bayesian Process (RABP) model is introduced as $RABP(G_0, \pi)$, where a base measure G_0 and a concentration parameter π are defined. The generative process for the RABP is provided below.

(1) Draw θ_1 from G_0

(2) For $t > 1$

(a) Sample ϵ from a Dirichlet distribution with parameter π, denoted by Dir(π), where $\epsilon = (\epsilon_1, ..., \epsilon_{c+1})^T$;

(b) Assuming ϵ_i is the probability of drawing θ^t from $\delta_{\theta(t-i)}$, where i is between 1 and C, and δ_a represents a discrete distribution with probability mass function equal to 1 at point a;

(c) The value of θ^t is drawn from G_0 with probability ϵ_{c+1}.

G_0 is the basis distribution in this generating process, and C is the length of the timeframe. The attention signal is defined as $\epsilon = (\epsilon_1, ..., \epsilon_{c+1})$, a $(C + 1)$-dimensional vector. The importance of a previous neighbor θ^{t-i} to θ^t is represented by it. The sum of ϵ from $i = 1$ to $C + 1$ equals 1, as it is distributed according to the Dirichlet distribution. Fig. 2.1b demonstrates the RABP graphical model. θ^t can be represented as follows, based on the generative process, where K is the length of each θ^i [37]:

$$\theta^t | \theta^{t-C:t-1}, G_0, \pi \sim \sum_{i=1}^{C} \epsilon_i \cdot \theta_{t-i} + \epsilon_{C+1} \cdot G_0 \tag{13}$$

The attention signal ϵ_i denotes the importance of the preceding parameter within the sequence.

While the Recurrent Chinese Restaurant Process (RCRP) takes into account both the position and content information, the Recurrent Attentional Bayesian Process (RABP) considers multiple previous time points with varying attentional weights. On the other hand, the RABP models recurrent sequences in discrete space with attention signals [37].

Bi-Directional Recurrent Attentional Bayesian Process

To model sequential data in both directions, the bi-RABP was developed as an extension of the RABP. In the bi-RABP, θ^t within $bi - RABP(G_0, \pi)$ can be expressed as follows [37]:

$$\theta | \theta^{t-c:t-1}, \theta^{t+c:t+1}, G_0, \pi \sim \sum_{i=1}^{C} \epsilon_i \cdot \theta^{t-i} + \sum_{j=1}^{c} \epsilon_j + c \cdot \theta^{t+j} + \epsilon_{2c+1} \cdot G_0 \tag{14}$$

The impact of the backward sequence on θ^t is calculated in the second term. In contrast to RABP, bi-RABP generates the current node by considering both the previous and the next nodes simultaneously. This approach could be advantageous for sequential data like text, as it allows for capturing local information from both directions [37]. Fig. 2.1c illustrates the bi-RABP graphical model.

Bi-Directional Recurrent Attentional Topic Model

In linguistics, coherence refers to how closely a sentence's subject distribution relates to sentences and subsequent sentences. The preceding phrases influence the sentence topics in the forward direction, which corresponds to the RABP's motivation. From the backward direction, the subsequent sentences, which can be represented by another RABP, have an impact on the sentence topic. Thus, a sentence's topic distribution is derived from its preceding and subsequent sentences. For each sentence, a simple bag-of-words language model is used in same way as the other models [13, 37].

θ^{sj} signifies the topic distribution over K latent topics of the j^{th} sentence $\boldsymbol{s}_j$ in document D. θ^{sj} is generated by $\theta^{t-C:t-1}, \theta^{t+1:t+C}$ and G_0 using Eq. 14 based on the bi-RABP. Therefore, the bi-RATM generating process is defined as follows:

(1) To generate each topic k, sample β_k from a Dirichlet distribution with parameter vector μ of dimension V.

(2) For each document $\boldsymbol{d}^i, i \in \{1, ..., M\}$

  (a) Draw $\nu^d \sim Dir(\alpha)$

  (b) For the sentence $\boldsymbol{s}_j, j \in \{1, ..., S_i\}$ in the document $\boldsymbol{d}$:

    i. Draw $\theta^{s_j} \sim bi - RABP(\delta_{\nu^d}, \pi)$

    ii. For each word $w_n, n \in \{1, ..., N_j\}$ in sentences $\boldsymbol{s}_j$:

      A. Draw $z_n \sim Mult(\theta^{s_j})$

      B. Draw $w_n \sim Mult(\beta_{z_n})$

2.2.4 Dirichlet Multinomial Regression

Several topic models consider the relationships between various documents. One such model is the relational topic model (RTM) [84], which represents the link between a pair of documents by using a binary random variable, based on the latent topic assignments of both documents[84, 32].

xLDA [85], on the other hand, is a variation of DMR-based models that incorporates relational information. It achieves this by employing a relational Gaussian process prior on document-specific Dirichlet parameters, thereby enabling the model to capture both metadata and document relations. This kernel-based method is highly adaptable, but it requires effort to make optimization scalable and to choose a sparse subset of specified relation types [32].

DMR encompasses a wide range of conditional topic models, which provide users with considerable flexibility to define novel features. To represent each document d, a vector, denoted as x_d, is constructed. This vector captures the relevant metadata values as features. To illustrate, if the presence or absence of authors is a metadata feature, then x_d will include a value of 1 in each position corresponding to the author(s) listed for document d, and a value of 0 in all other positions. Additionally, to consider the mean value of each topic, a default feature is included in the model, which always has a value of 1 [67]. A vector, denoted as λ_t, is assigned to each topic t, with the length of the vector being equal to the number of features. The generative process for the feature matrix X is as follows, where X represents the data matrix [67].

Firstly, for each topic, denoted as t, we randomly draw from a normal distribution $\mathcal{N}(0, \sigma^2 I)$, assigning the result to λ_t, and from a Dirichlet distribution $\mathcal{D}(\beta)$, assigning the outcome to ϕ_t. Then, for each document d, we compute α_{dt} for every topic t, out of a total of T topics, as the exponential of the dot product of the document's metadata vector x_d and λ_t. Next, we make a random draw from a Dirichlet distribution $\mathcal{D}(\alpha_d)$, assigning the result to θ_d, which symbolizes the document's topic distributions. Lastly, for each word i in the document, we draw randomly from a multinomial distribution $\mathcal{M}(\theta_d)$, assigning the outcome to z_i (the topic allocated to the word), and from a multinomial distribution $\mathcal{M}(\phi_{z_i})$, assigning the result to w_i (the actual word generated given the topic z_i). This procedure forms the cornerstone of various topic modeling techniques, essential in discovering the latent themes in a set of documents.

The model involves three fixed parameters, namely, σ^2, which represents the variance of the prior on parameter values, β, which is the Dirichlet prior on the topic-word distributions, and $|T|$. The model is trained using a stochastic EM sampling technique, which involves alternating between sampling topic assignments from the current prior distribution conditioned on the observed words and features, and numerically optimizing the parameters based on the topic assignments [67]. To

implement this, the standard L-BFGS optimizer [86] and Gibbs sampling–based LDA trainer are used [67].

2.2.5 Deep Dirichlet Multinomial Regression

The dDMR (deep Dirichlet multinomial regression) model is an extension of DMR. It incorporates a deep neural network to convert input metadata into features, which are then utilized to form the Dirichlet hyperparameter. DMR uses a log-linear function to parameterize document-topic priors based on document features. In contrast, dDMR jointly learns a feature representation for each document and a log-linear function that captures the distribution over topics. As the neural network maps document features to topic priors, the topic model and neural network parameters are jointly optimized using gradient ascent and back-propagation [71, 67].

In dDMR, the log-linear model is replaced by an arbitrary function that maps a real-valued vector of a certain dimension to another representation of a different dimension. The preference for neural networks stems from their powerful expressive capacity, their proficiency in effectively generalizing to novel data, and their unique capability for joint training.

Dirichlet multinomial regression (DMR) and dDMR topic modeling are powerful tools in natural language processing for analyzing text data. DMR models the relationship between a set of predictor variables and a set of categorical response variables using a multinomial distribution with a Dirichlet prior. dDMR extends this approach by introducing a neural network layer to capture nonlinear relationships between predictor and response variables [71, 67].

Inspired by the success of DMR and dDMR, as well as the need for more flexible and robust approaches, we propose a novel extension that leverages the GD and Beta-Liouville distributions to further enhance the model's performance in capturing complex distributions in data.

2.2.6 BERTopic Embedding

In BERTopic, the process begins with generating document embeddings using pre-trained transformer-based language models, such as SBERT [74]. SBERT transforms sentences and documents into dense vector representations that capture their semantic content, positioning semantically similar texts close to each other in the vector space, which is essential for effective clustering. This step

leverages SBERT's Siamese network structure to produce embeddings optimized for semantic similarity tasks. While SBERT is commonly used, BERTopic's design allows for the integration of other embedding techniques, ensuring adaptability to advancements in embedding technologies [24]. Following the generation of high-dimensional embeddings, the next step involves reducing their dimensionality to facilitate efficient clustering. This is achieved using Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique that preserves both local and global structures of the data [75]. UMAP projects the high-dimensional embeddings into a lower-dimensional space, making the clustering process more effective and computationally feasible by addressing the challenges posed by the high-dimensional space, such as the curse of dimensionality. This combination of generating rich, contextual embeddings and reducing their dimensionality ensures that BERTopic can produce coherent and contextually relevant topics, leveraging the latest developments in natural language processing while maintaining efficient and accurate clustering capabilities [24].

2.2.7 Multi-grain clustering topic model

The Multi-Grain Clustering Topic Model (MGCTM) is a sophisticated framework that integrates document clustering and topic modeling into a cohesive system, enhancing the performance of both tasks. This model operates under the assumption that a document corpus can be divided into several latent groups, each characterized by unique local topics and shared global topics. Local topics are specific to each group, capturing the distinct semantics of the documents within that group, while global topics represent common themes that span across all groups in the corpus [27].

MGCTM's generative process begins by assigning each document to a latent group based on a multinomial distribution parameterized by a prior probability vector π. For each document, two sets of topic proportions are sampled: local topic proportions $\theta^{(l)}$ from a group-specific Dirichlet prior $\alpha^{(l)}$ and global topic proportions $\theta^{(g)}$ from a global Dirichlet prior $\alpha^{(g)}$. To generate words in a document, a Bernoulli variable δ determines whether each word is generated from a local or global topic. If δ equals 1, the word is drawn from a local topic distribution $\beta^{(l)}$; if δ equals 0, the word is drawn from a global topic distribution $\beta^{(g)}$ [27].

To approximate the posterior distribution of latent variables, MGCTM employs variational inference, which is necessary due to the intractability of exact inference. The model defines a variational distribution to minimize the Kullback-Leibler (KL) divergence between the true posterior and the variational distribution. During the Expectation Step (E-Step), the model updates variational parameters by maximizing the lower bound of the data likelihood while keeping the model parameters fixed. In the Maximization Step (M-Step), the model parameters are optimized by maximizing the lower bound with fixed variational parameters.

Key parameter updates in MGCTM include adjusting the multinomial parameters (ζ) to reflect the probability of a document belonging to each group, refining the Dirichlet parameters ($\mu^{(l)}$, $\mu^{(g)}$) for local and global topic proportions, and optimizing the Bernoulli parameters (τ) to decide between local and global topic generation for words. The topic distributions ($\phi^{(l)}$, $\phi^{(g)}$) are also refined to improve the assignment of words to topics, enhancing the model's accuracy [27].

By jointly modeling clustering and topic extraction, MGCTM achieves more coherent and discriminative topic representations. This integrated approach distinguishes between fine-grained local details and broad global themes, significantly improving the meaningfulness of document clusters. The model's ability to leverage topic information for clustering results in more precise and relevant topic representations, making it a powerful tool for analyzing complex document corpora.

Chapter 3

Hidden variable models in text classification and sentiment analysis

3.1 Introduction

Digital data has greatly improved information gathering across many fields in today's fast-paced technological world. Effective and scalable statistical models are needed to uncover hidden insights in massive amounts of digital and textual data created daily [1].

Analyzing textual data is difficult due to its complexity. This issue has been addressed in various ways. The rapid advancement of information technology has made massive amounts of scientific documents available for mining, making their analysis an increasingly popular subject for study.

Data projection and clustering are crucial for document analysis, with projection aimed at creating low-dimensional, meaningful data representations and clustering grouping similar data patterns [3, 4]. Traditionally, these methods are studied separately, but they intersect in many applications [4]. K-means clustering, though widely used for creating compact cluster representations, does not fully capture document semantics. This gap has led to the adoption of machine learning and deep learning for text mining challenges, including text classification [5], summarization [6], segmentation [7], topic modeling [8], and sentiment analysis [9].

In this chapter, we will focus on topic modeling aspects specifically matrix decomposition-based and generative models [10]. PLSA and other matrix decomposition methods mining text require

corpus structure knowledge to identify relevant words for each topic [11, 12]. Using latent variables, generative models like LDA analyze word context and view documents as topic mixtures [4]. LDA enhances PLSA with Dirichlet distributions for enhanced generative capabilities [13]. However, current models often lack clear probability models with hidden variables [12, 16, 17, 18]. Models such as variational EM in MPCA, extensions of LDA, Gibbs Sampling, and MCMC methods have been developed to address this issue, but they require longer convergence times [22].

In this chapter, we introduce two novel models, GDMPCA and BLMPCA, that significantly improve text classification and sentiment analysis by combining Generalized Dirichlet (GD) and Beta-Liouville (BL) distributions for a more in-depth understanding of text data complexities [87, 88, 89]. Both models employ variational Bayesian inference and collapsed Gibbs sampling for efficient and scalable computational performance which is critical for handling large datasets.

The Generalized Dirichlet (GD) distribution, introduced by [90], exhibits a more flexible covariance structure than its Dirichlet counterpart. Similarly, the Beta-Liouville (BL) distribution, enriched with additional parameters, offers improved adjustments for data spread and modeling efficiency. Our contribution is validated by rigorous empirical evaluation on real-world datasets, which demonstrates our models' superior accuracy and adaptability. This work represents a significant step forward in text analysis methodologies, bridging theoretical innovation with practical application, with experimental results demonstrating the relationships between these models.

3.2 Models

In this section, we present two pioneering models, Generalized Dirichlet Multinomial Principal Component Analysis (GDMPCA) and Beta-Liouville Multinomial Principal Component Analysis (BLMPCA), designed to revolutionize text classification and sentiment analysis. At the core of our approach is the integration of GD and BL distributions, respectively, into the PCA framework. This integration is pivotal, as it allows for a more nuanced representation of text data, capturing the inherent sparsity and thematic structures more effectively than traditional methods.

The GDMPCA model leverages the flexibility of the Generalized Dirichlet distribution to model

the variability and co-occurrence of terms within documents, enhancing the model's ability to discern subtle thematic differences. On the other hand, the BLMPCA model utilizes the Beta-Liouville distribution to precisely capture the polytopic nature of texts, facilitating a deeper understanding of sentiment and thematic distributions. Both models employ variational Bayesian inference, offering a robust mathematical framework that significantly improves computational efficiency and scalability. This approach not only aids in handling large datasets with ease but also ensures that the models remain computationally viable without sacrificing accuracy.

To elucidate the architecture of our proposed models, we delve into the algorithmic underpinnings, detailing the iterative processes that underlie the variational Bayesian inference technique. This includes a comprehensive discussion of the optimization strategies employed to enhance convergence rates and ensure the stability of the models across varied datasets. Moreover, we provide a comparative analysis, drawing parallels and highlighting distinctions between our models and existing text analysis methodologies. This comparison underscores the superior performance of GDMPCA and BLMPCA in terms of accuracy, adaptability, and computational efficiency, as evidenced by extensive empirical evaluation on diverse real-world datasets.

Our exposition on the practical implications of these models reveals their broad applicability across numerous domains, from automated content categorization to nuanced sentiment analysis in social media texts. The innovative aspects of the GDMPCA and BLMPCA models, coupled with their empirical validation, underscore their potential to set a new standard in text analysis, offering researchers and practitioners alike powerful tools for uncovering insights from textual data.

Table 3.1 summarizes the relevant variables for the proposed models.

Table 3.1: Parameters of Generalized Dirichlet and Beta-Liouville Distributions

Parameter	Generalized Dirichlet (GDMPCA)	Beta-Liouville (BLMPCA)
ξ	Parameters of GD distribution	Not applicable
Υ	Not applicable	Parameters of BL distribution
m	Mixture weights (GD)	Mixture weights (BL)
z	Topic assignments	Topic assignments
w	Words in documents	Words in documents
Ω	Multinomial parameters (words)	Multinomial parameters (words)
L	Count of words present in a document	Count of words present in a document
C, Ω_k, c_k	Multinomial parameters for topics	Multinomial parameters for topics

3.3 Generalized Dirichlet Multinomial PCA

Bouguila [91] demonstrated that when mixture models are used, GD distribution is a reasonable alternative to the Dirichlet distribution for clustering count data.

The GD distribution, like the Dirichlet distribution, is a conjugate prior for the multinomial distribution, as mentioned above. In addition, the GD distribution includes a covariance matrix that is more broad [91].

Hence, the variational Bayes method will be employed to create an expansion of the MPCA model that includes the GD assumption. The GDMPCA is expected to function efficiently due to the fact that the Dirichlet distribution is a particular example of the GD [92]. GDMPCA, similar to MPCA, is a comprehensive generative model that is utilized on a corpus. The corpus, defined by $D = \{w_1, w_2, \ldots, w_M\}$, represents a collection of M documents. Each document, denoted as w_m, is comprised of a sequence of N_m words. The words in a document are represented by binary vectors from a vocabulary of V words. If the j-th word is selected, it is represented by $w_j^n = 1$, otherwise it is represented by $w_j^n = 0$ [46]. The GDMPCA model thereafter delineates the production of each word in the document via a sequence of stages that encompass c, a $d + 1$ dimensional binary vector of topics:

$$
\begin{aligned}
m &\sim GD(\xi) \\
z &\sim Multinomial(m, L) \\
w_k &\sim Multinomial(\Omega_k, c_k)
\end{aligned}
\tag{15}
$$

If the i-th topic is chosen $z_i^n = 1$ in other cases $z_i^n = 0$. $m = (m_1, \ldots, m_{d+1})$, where $m_{d+1} = 1 - \sum_{i=1}^{d} m_i$.

The multinomial probability $p(w_n \mid z_n, \Omega_w)$ is conditioned on the variable z_n. The distribution $GD(\xi)$ is a d-variate Generalized Dirichlet distribution characterized by the parameter set $\xi = (a_1, b_1, \ldots, a_d, b_d)$, with its probability distribution function denoted by p, where $\eta_i = b_i - a_{i+1} - b_{i+1}$ [46]:

$$
p(m_1, \ldots, m_d|\xi) = \prod_{i=1}^{d} \frac{\eta(a_l + b_l)}{\eta(a_l)\eta(b_l)} m_i^{a_l - 1} (1 - \sum_{j=1}^{i} m_j)^{\eta_i} \tag{16}
$$

The GD distribution simplifies to a Dirichlet distribution when $b_i = a_{(i+1)} + b_{(i+1)}$.

The mean, variance and the covariance matrix of the GD distribution are as follows [92] :

$$
E(m_i) = \frac{a_l}{a_l + b_l} \prod_{k=1}^{i-1} \frac{b_k}{a_k + b_k} \tag{17}
$$

$$
var(m_i) = E(m_i) \left(\frac{a_l + 1}{a_l + b_l + 1} \prod_{k=1}^{i-1} \frac{b_k + 1}{a_k + b_k} + 1 - E(\theta_i) \right) \tag{18}
$$

and the covariance between m_i and m_j is given by :

$$
cov(m_i, m_j) = E(m_j) \left(\frac{a_l}{a_l + b_l + 1} \prod_{k=1}^{i-1} \frac{b_k + 1}{a_k + b_k} + 1 - E(m_i) \right) \tag{19}
$$

The covariance matrix of the GD distribution offers greater flexibility compared to the Dirichlet distribution, due to its more general structure. This additional complexity allows for an extra set of parameters, providing $d - 1$ additional degrees of freedom, which enables the GD distribution to more accurately model real-world data. Indeed, the GD distribution fits count data better than

the commonly used Dirichlet distribution [93]. The Dirichlet and GD distributions are both members of the exponential family (Appendix 8.1). Furthermore, they are also conjugate priors to the multinomial distribution.

As a result, we can use the following method to learn the model.

The likelihood for the GDMPCA is given as follows:

$$p(m, w|\xi, \Omega) = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} z^{L}_{w_{1,1}, w_{1,2}..., w_{k,1}, w_{1,J}...w_{K,J}} m_k^{b_{k-1}-1} \prod_{i=1}^{k-1} [m_i^{a_{i-1}} \\ (\sum_{j=1}^{k} m_j)^{b_{i-1}+(a_i+b_i)}] \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \tag{20}$$

Hence, when hidden variables are assigned GD priors, and given a defined universe of words, we use an empirical prior derived from the observed proportions of words in the universe, denoted by f, where $\sum_k f_k = 1$. The equation then is structured as follows:

$$m \sim GD(\xi) \\ \Omega_k \sim GD(2f) \tag{21}$$

where 2 shows the small size of the prior sample size.

First, we will calculate the parameters of GD by utilizing the Hessian matrix as stated in Appendix 8.2.1, by utilizing equations 20 and 21. In order to determine the most effective variational parameters, we want to minimize the KL divergence between the variational distribution and the posterior distributions $p(m, w|\Omega, \xi)$. This is accomplished using an iterative fixed-point algorithm. The variational parameters are specified as follows:

$$q(m, c|\eta, \Phi) = q(m|\eta) \prod_{k=1}^{K} q(c_k|\Phi_k) \tag{22}$$

As an alternative to the posterior distribution $p(m, c, w, \xi, \Omega)$, we determine the variational parameters η and Φ through a detailed optimization process outlined subsequently. To simplify, Jensen's inequality is applied to establish a lower bound on the log-likelihood, which allows us to

disregard the parameters η and Φ [94]:

$$\begin{aligned}
\log p(w|\xi,\Omega) &= \log \int \sum_z p(m,c,w|\xi,\Omega)dm \\
&= \log \int \sum_z \frac{p(m,c,w|\xi,\Omega)q(m,c)}{q(m,c)}dm \\
&\geq \int \sum_z \log p(m,c,w|\xi,\Omega)q(m,c)dm \\
&- \int \sum_z q(m,c)\log q(m,c)dm \\
&= E[\log p(m,c,w|\xi,\Omega)] - E[\log q(m,c)]
\end{aligned} \tag{23}$$

Consequently, Jensen's inequality establishes a minimum value for the log likelihood of a specific variational distribution $q(m,c|\eta,\Phi)$.

The expression $\mathcal{L}(\eta,\Phi;\xi,\Omega)$ in Eq. 23 represents the second side of the equation. The difference between both sides of this equation corresponds to the KL divergence between the variational distribution and the real posterior probability. This reaffirms the significance of the variational variables, resulting in the subsequent equation:

$$\log p(w|\xi,\Omega) = \mathcal{L}(\eta,\Phi;\xi,\Omega) + D(q(m,c|\eta,\Phi))|p(m,c|x,\xi,\Omega) \tag{24}$$

As shown in Equation 24, the process of maximizing the lower bound $\mathcal{L}(\eta,\Phi;\xi,\Omega)$ with respect to η and Φ is analogous to reducing the Kullback-Leibler (KL) divergence between the variational posterior probability. A lower limit can be described by factorizing the variational distributions:

$$\begin{aligned}
\mathcal{L}(\eta,\Phi;\xi,\Omega) &= E_q[\log p(m|\xi)] + E_q[\log p(c|m)] + E_q[\log p(w|c,\Omega)] \\
&- E_q[\log q(m)] - E_q[\log q(c)]
\end{aligned} \tag{25}$$

After that, we can extend Eq. 25 in terms of the model parameters (ξ, Ω) and variational parameters (η, Φ).

$$\begin{aligned}
\mathcal{L}(\eta, \Phi; \xi, \Omega) = & \sum_{l=1}^{d}[\log \eta(a_l + b_l) - \log \eta(a_l) - \log \eta(b_l)] \\
& + \sum_{l=1}^{d}[a_l(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) \\
& + (\Psi(\Phi) - \Psi(\eta_l + \Phi))(a_l - a_{l+1} - b_{l+1})] \\
& + \sum_{n=1}^{N}\sum_{l=1}^{d} m_{nl}(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + \\
& \sum_{n=1}^{N} m_{n(d+1)}(\Psi(\Phi) - \Psi(\Phi + \eta_d)) \\
& + \sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{v} m_{nl} w_n^j \log(\Omega_{ij}) \\
& - \sum_{l=1}^{d}(\log \eta(\eta_l + \Phi) \log \eta(\eta_l) - \log \eta(\Phi)) \\
& - \sum_{l=1}^{d}[\eta_l(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + (\Psi(\Phi) - \Psi(\Phi + \eta_l)) \\
& (\Phi - \eta_{l+1} - \Phi_{l+1})]
\end{aligned} \tag{26}$$

In order to find ϱ_{nl}, we proceed to maximize with the respect to ϱ_{nl}, so we have the following equations:

$$\begin{aligned}
\mathcal{L}[m_{nl}] = & m_{nl}(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + m_{nl} \log \Omega_{w(lv)} - m_{nl} \log m_{nl} \\
& + \tau_n(\sum_{ll=1}^{d+1} m_{n(ll)} - 1)
\end{aligned} \tag{27}$$

Consequently, we include the following:

$$\frac{\partial \mathcal{L}}{\partial \varrho_{nl}} = (\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + \log \Omega_{lv} - \log \varrho_{nl} - 1 + \tau_n \tag{28}$$

Setting the aforementioned equation to zero results in:

$$m_{nl} = \Omega_{lv} e^{(\tau_n - 1)} e^{(\Psi(\eta_l) - \Psi(\eta_l + \Phi))} \tag{29}$$

Following that, we enhance Eq. 26 with regard to η_i. The terms containing η_i are:

$$\begin{aligned}
\mathcal{L}[\xi_q] = & \sum_{l=1}^{d} [a_l(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + (\Psi(\eta_l) - \\
& \Psi(\eta_l + \Phi))(b_l - a_{l+1} - b_{l+1})] \\
& + \sum_{n=1}^{N} m_{nl}(\Psi(\eta_l) - \Psi(\eta_l + \Phi) + \sum_{n=1}^{N} m_{n(d+1)}(\Psi(\eta_d) - \Psi(\eta_d + \Phi_d)) \\
& - \sum_{l=1}^{d} (\log \eta(\eta_l + \Phi_l) - \log \eta(\eta_l) - \log \eta(\Phi_l)) \\
& + \sum_{l=1}^{d} (\Psi(\eta_l) - \eta_l(\Psi(\eta_l + \Phi_l)) \\
& + (\Psi(\Phi) - \Psi(\Phi + \eta_l))(\Phi - \eta_{l+1} - \Phi_{l+1})))
\end{aligned} \tag{30}$$

By equating the derivative of the given equation to zero, we obtain the following updated parameters:

$$\eta_l = a_l + \sum_{n=1}^{N} m_{nl} \tag{31}$$

$$\Phi_l = b_l + \sum_{n=1}^{N} \sum_{l=l+1}^{d+1} m_{n(l)} \tag{32}$$

The challenge of obtaining empirical Bayes estimates for model parameters ξ and Ω entails utilizing the variational lower bound as an approximation for the marginal log probability, employing variational parameters η and Φ. The estimations are derived by optimizing this lower bound with respect to ξ and Ω. Previously, our attention was directed towards the log likelihood of an individual document. The total variational lower limit is obtained by adding all the lower bounds from each document. In the M-step, we aim to maximize the bound for the parameters ξ and Ω, which is similar to conducting coordinate ascent as seen in Equation 33. The equation for updating Ω is derived by isolating terms and applying Lagrange multipliers to optimize the constraint related to Ω.

$$\mathcal{L}[\Omega] = \sum_{d=1}^{M}\sum_{n=1}^{N_s}\sum_{l=1}^{K+1}\sum_{j=1}^{V} m_{dnl} w_{dn}^{j} \log \Omega_{(lj)} + \sum_{l=1}^{K+1} \tau_l \big(\sum_{j=1}^{V} \Omega_{w(ij)} \big) \tag{33}$$

To derive the update equation for $\Omega_{(lj)}$, we compute the slope of the variational lower limit with regard to $\Omega_{(lj)}$ and set this derivative to zero. This step guarantees that we identify the exact point at which the lower limit reaches its maximum value with regard to the parameter $\Omega_{(lj)}$.

$$\Omega_{(lj)} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j} \tag{34}$$

The updates mentioned lead to convergence at a local maximum of the $\log p(\Omega, \xi | r)$, which is the most efficient choice for all product approximations of the form $q(m)q(w)$ for the joint probability $p(m, w|\Omega, \xi, r)$. This approach ensures that the variational parameters are adjusted to optimally approximate the true posterior distributions within the constraints of the model.

$$\Phi_l = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} m_{nl} (\Psi(\eta_l) - \Psi(\eta_l + \Phi)) \tag{35}$$

$$\eta_l = a_l + \sum_{n=1}^{N} m_{nl} \tag{36}$$

$$\Omega_{(lj)} = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} (2 f_j \sum_{d=1}^{M}\sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j}) \tag{37}$$

3.3.1 Collapsed Gibbs Sampling Method

Utilizing the fundamental procedure of the GD distribution as delineated in the all-encompassing generative formula $p(c, z, \theta, \varphi, w|, \Omega, \xi, \mu)$ within our innovative methodology, we can express it in the following manner:

$$p(c, z, \theta, \varphi, w|, \Omega, \xi, \mu) = p(w|\mu) p(\theta|\Omega) p(\varphi|\xi) \times \prod_{n=1}^{N} p(z_n|\theta) p(x_n|z_{nn}, \varphi) \tag{38}$$

Here, $p(\theta|\Omega)$ signifies the GD document prior distribution, where $\Omega = (a_1, b_1, \ldots, a_n, b_n)$ serves as the hyperparameter. Simultaneously, $p(\varphi|\xi)$, with $\xi = (\alpha_1, \beta_1, \ldots, \alpha_d, \beta_d)$ as its hyperparameters, represents the GD corpus prior distribution. The process of Bayesian inference seeks

to approximate the posterior distribution of hidden variables z by integrating out parameters, which can be mathematically depicted as:

$$p(c, z|w, \Omega, \xi) = W \int_{\theta} \int_{\varphi} p(c, z, \theta, \varphi, |\Omega, \xi) d\varphi d\theta \tag{39}$$

Crucially, the joint distribution is expressed as a product of Gamma functions, as highlighted in prior research [13, 95, 64]. This expression facilitates the determination of the expectation value for the accurate posterior distribution:

$$p(z_{ij} = k|c, w, \Omega, \xi) = \mathrm{E}_{p(z^{-ij}|w,c,\Omega,\xi)}[p(z_{ij} = k|z^{-ij}, c, w, \Omega, \xi)] \tag{40}$$

Employing the GD prior results in the posterior calculation as outlined below:

$$\begin{aligned} p(z_{ij} = k|z^{-ij}, c, w, \Omega, \xi) \propto & \left[\frac{(N_{jk}^{-ij} + \alpha_{wk})(\beta_{wk} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})}{(\alpha_{wk}\beta_{wk} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})} \right] \\ & \times \left[\frac{(N_{kv_{ij}}^{-ij} + a_v)(b_v + \sum_{d=v}^{V+1} N_{kd_{ij}}^{-ij})}{a_v + b_v + \sum_{d=v}^{V+1} N_{kd_{ij}}^{-ij})} \right] = A(K) \end{aligned} \tag{41}$$

This leads to a posterior probability normalization as shown:

$$p(z_{ij} = k|z^{-ij}, x, \Omega, \xi) = \frac{A(k)}{\sum_{k'=1}^{K} A(k')} \tag{42}$$

The sequence from Eq. 40 to Eq. 42 delineates the complete collapsed Gibbs sampling procedure, encapsulated as:

$$p(z_{ij} = k|c, w, \Omega, \xi) = \mathrm{E}_{p(z^{-ij}|w,c,\Omega,\xi)} \left[\frac{A(k)}{\sum_{k'=1}^{K} A(k')} \right] \tag{43}$$

The implementation of collapsed Gibbs sampling in our GD-centric model facilitates sampling directly from the actual posterior distribution p, as indicated in Eq. 43. This sampling technique is deemed more accurate than those employed in variational inference models, which typically approximate the distribution from which samples are drawn [96, 64]. Hence, our model's precision

is ostensibly superior.

Upon completion of the sampling phase, parameter estimation is conducted using methodologies discussed in [96, 97]

3.4 Beta-Liouville Multinomial PCA

For Beta-Liouville Multinomial PCA (BLMPCA) model, we define a corpus as a collection of documents with the same assumption described in the GDMPCA section. Hence, we have the following procedure for the model on every single word of the document. The BLMPCA model generates each word in the document using the following processes, where c is a binary vector of topics with $d + 1$ dimensions:

$$\begin{aligned} m &\sim BL(\Upsilon) \\ z &\sim Multinomial(m, L), \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{44}$$

The model described utilizes binary variables to represent topics for each word, where a chosen topic is indicated by $z_i^n = 1$ and not chosen by $z_i^n = 0$. The vector z_n represents topic assignments across all topics for a word. Another vector, m, captures the distribution of topic proportions across the document. Each chosen topic is associated with a multinomial prior over the vocabulary, allowing words to be drawn randomly based on assigned topics. The probability $p(w_n|z_n, \Omega_w)$ models the likelihood of each word given topic assignments and topic-word distributions [98].

Additionally, $BL(\Upsilon)$ represents a d-variate BL distribution with parameters $\Upsilon = (\kappa_1, ..., \kappa_D, \kappa, \iota)$. The probability distribution function of this Beta-Liouville distribution encapsulates the prior beliefs about the distribution of topics across documents, accommodating complex dependencies among topics and allowing for flexibility in modeling topic prevalence and co-occurrence within the corpus.

$$P(\theta_1,\ldots,\theta_D|\Upsilon) = \frac{\eta(\sum_{d=1}^{D}\kappa_d)\eta(\kappa+\iota)}{\eta(\kappa)\eta(\iota)}\prod_{d=1}^{D}\frac{\theta_d^{\kappa_d-1}}{\eta(\kappa_d)}\times \left(\sum_{d=1}^{D}\theta_d\right)^{\kappa-\sum_{l=1}^{D}\kappa_l}\times\left(1-\sum_{l=1}^{D}\theta_l\right)^{\iota-1} \tag{45}$$

Dirichlet distribution is the special case of BL if $\iota_d = \kappa_{d+1} + \iota_{d+1}$ [**?**, 46].

The statistical measures of the mean, variance, and covariance for the BL distribution are [95]:

$$E(\theta_d) = \frac{\kappa}{\kappa+\iota}\frac{\kappa_d}{\sum_{d=1}^{D}\kappa_d} \tag{46}$$

$$\begin{aligned} var(\theta_d) = (\frac{\kappa}{\kappa+\iota})^2\frac{\kappa_d(\kappa_d+1)}{(\sum_{m=1}^{D}\kappa_m)(\sum_{m=1}^{D}\kappa_m+1)} \\ - E(\theta_d)^2\frac{\kappa_d^2}{(\sum_{m=1}^{D}\kappa_m)^2} \end{aligned} \tag{47}$$

The covariance between θ_l and θ_k is determined by:

$$Cov(\theta_l,\theta_k) = \frac{\kappa_l\kappa_k}{\sum_{d=1}^{D}\kappa_d}\left(\frac{\frac{(\kappa+1)(\kappa)}{(\kappa+\iota+1)(\kappa+\iota)}}{\sum_{d=1}^{D}\kappa_d+1} - \frac{\frac{\kappa}{\kappa+\iota}}{\sum_{d=1}^{D}\kappa_d}\right) \tag{48}$$

The earlier equation illustrates that the covariance matrix of the BL distribution offers a broader scope compared to the covariance matrix of the Dirichlet distribution. For the parameter estimation of BLMPCA first the parameter Ω was estimated by the Beta Liouville prior on m using parameters Υ [18]. The likelihood model for the BLMPCA is given as follows:

$$\begin{aligned} p(m,w|\Upsilon,\Omega) = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D}\kappa_d)\eta(\kappa+\iota)} z^{L}_{w_{1,1},w_{1,2}\ldots,w_{k,1},w_{1,J}\ldots w_{K,J}}\Big[\frac{1}{\eta(\kappa_d)}m_k^{\kappa_d-1}+ \\ \sum_k m_k^{\kappa-\sum_d\kappa_d} + (1-\sum_k m_k)^{\iota-1}\Big]\prod_{k,j} m_k^{w_{k,j}}\Omega_{k,j}^{w_{k,j}} \end{aligned} \tag{49}$$

For the BL priors, we have the following:

$$\begin{aligned} m &\sim BL(\Upsilon) \\ \Omega_k &\sim BL(2f) \end{aligned} \tag{50}$$

In the following step, we will estimate the parameters for Ω using the Beta Liouville prior and the Hessian matrix (Appendix 8.3). As we explained in the previous section 3.3, we should estimate the model parameters (Υ, Ω) and the variational parameters (η, Φ), according to Eqs. 22, 23 and 25 to find m_{nl} and we proceed to maximize with the respect to m_{nl} so we have following equations:

$$
\begin{aligned}
\mathcal{L}(\eta, \Phi; \Upsilon, \Omega) &= \log(\eta(\sum_{d=1}^{D} \kappa_d)) + \log(\eta(\kappa + \iota)) - \log(\eta(\kappa)) \\
&- \log(\eta(\iota)) - \sum_{d=1}^{D} \log \eta(\kappa_d) + \sum_{d=1}^{D} \kappa_d (\Psi(\eta_d) - \Psi(\sum_{l=1}^{D} \eta_l) \\
&+ \kappa(\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) + \iota(\Psi(\iota_\eta) \\
&- \Psi(\kappa_\eta + \iota_\eta)) + \iota(\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\
&+ \sum_{n=1}^{N} \sum_{d=1}^{D} m_{nd} (\Psi(\eta_d) - \Psi(\sum_{l=1}^{D} \eta_l) + \Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\
&+ \sum_{n=1}^{N} m_{n(D+1)} (\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\
&+ \sum_{n=1}^{N} \sum_{l=1}^{D+1} \sum_{j=1}^{V} m_{nl} w_n^j \log(\Omega_{lj}) \\
&- \big(\log(\eta(\sum_{l=1}^{D} \kappa_l)) + \log(\eta(\kappa + \iota)) - \log \eta(\kappa) - \log \eta(\iota) \\
&- \sum_{i=1}^{D} \log \eta(\kappa_i) \\
&+ \sum_{i=1}^{D} \kappa_i (\Psi(\eta_{mi}) - \Psi(\sum_{l=1}^{D} \eta_{m(l)})) + \kappa_(\Psi(\kappa_{m\eta}) \\
&- \Psi(\kappa_{m\eta} \iota_{m\eta})) + \iota(\Psi(\iota_{m\eta}) - \Psi(\kappa_{m\eta} + \iota_{m\eta}))\big) \\
&- \big(\sum_{n=1}^{N} \sum_{l=1}^{D+1} m_{nl} \log(m_{nl}) \big)
\end{aligned}
\tag{51}
$$

To find m_{nl} , we proceed to maximize with respect to ϱ_{nl}:

$$\begin{aligned}\mathcal{L}[m_{nl}] = m_{nl}(\Psi(\eta_i) - \Psi(\sum_{l=1}^{D}\eta_l)) + m_{nl}\log\iota_{w(iv)} - m_{nl}\log(m_{nl}) \\ + \tau_n(\sum_{l=1}^{D} m_{nl} - 1)\end{aligned} \tag{52}$$

Therefor we have:

$$\frac{\partial\mathcal{L}}{\partial\varrho_{nl}} = (\Psi(\eta_d) - \Psi(\sum_{l=1}^{D}\eta_l)) + \log\iota_{w(iv)} - \log\varrho_{nl} - 1 + \tau_n \tag{53}$$

The next step is to optimize Eq. 51, to find the update equations for the variational; we separate the terms containing the variational Beta-Liouville parameters once more.

$$\begin{aligned}\mathcal{L}[\xi_q] = \kappa_d(\Psi(\eta_d)) - \Psi(\sum_{l=1}^{D}\eta_l) + \kappa(\Psi(\kappa_\eta) - \Psi(\kappa_\eta \\ + \iota_\eta)) + \iota(\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\ + \sum_{n=1}^{N}\varrho_n(\Psi(\eta_l) - \Psi(\sum_{l=1}^{D}\eta_l) + \Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\ + \sum_{n=1}^{N}\varrho_{n(D+1)}(\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\ - (\log(\eta(\sum_{l=1}^{D}\eta_l)) + \log(\eta(\kappa_\eta + \iota_\eta) - \log(\eta(\kappa_\eta)) \\ - \log(\eta(\iota_\eta)) - \log(\eta(\eta_l)) \\ + \eta_l(\Psi(\eta_l) + \Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) - \Psi(\sum_{l=1}^{D}\eta_l) \\ + \kappa_\eta(\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\ + \iota_\eta(\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta)))\end{aligned} \tag{54}$$

By choosing the terms that contain variational BL parameters $\eta_i, \kappa_\eta, \iota_\eta$, we obtain:

$$\begin{aligned}\mathcal{L}(\eta_i) = \kappa_i(\Psi(\eta_i)) - (\sum_{l=1}^{D}\kappa_l)(\Psi(\sum_{l=1}^{D}\eta_l)) + \sum_{n=1}^{N}\varrho_{ni}(\Psi(\eta_i) - \Psi(\sum_{l=1}^{D}\eta_l)) \\ - (\log\eta(\sum_{l=1}^{D}) - \log\eta(\eta_i) + \eta_i(\Psi(\sum_{l=1}^{D}\eta_l)\sum_{d=1}^{D}\eta_d)\end{aligned} \tag{55}$$

and

$$\begin{aligned}\mathcal{L}[\kappa_\eta] = \kappa(\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) + \iota(-\Psi(\kappa_\eta + \iota_\eta)) \\ + (\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta))\sum_{n=1}^{N}\sum_{i=1}^{D}\varrho_{ni}\sum_{n=1}^{N}\varrho_{n(D+1)}(-\Psi(\kappa_\eta + \iota_\eta)) \\ - (\log(\kappa_\eta + \iota_\eta) - \log(\eta(\kappa_\eta)) + \kappa_\eta(\Psi(\kappa_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \\ + \iota_\eta(-\Psi(\kappa_\eta + \iota_\eta)))\end{aligned} \tag{56}$$

Setting Eqs. 54, 55 and 56 to zero we have the following update parameters:

$$\eta_i = \kappa + \sum_{n=1}^{N}\varrho_{ni} \tag{57}$$

$$\kappa_\eta = \kappa + \sum_{n=1}^{N}\sum_{d=1}^{D}\varrho_{nd} \tag{58}$$

$$\iota_\eta = \iota + \sum_{n=1}^{N}\varrho_{n(D+1)} \tag{59}$$

We address the challenge of deriving empirical Bayes estimates for the model parameters Υ and Ω by utilizing the variational lower bound as a substitute for the marginal log likelihood. This approach fixes the variational parameters η and Φ at values determined through variational inference. We then optimize this lower bound to achieve empirical Bayes estimates of the model parameters.

To estimate Ω_w, we formulate necessary update equations. The process of maximizing Eq. 54 with respect to Ω results in the following equation:

$$\mathcal{L}[\Omega_w] = \sum_{d=1}^{M}\sum_{n=1}^{N_s}\sum_{l=1}^{D+1}\sum_{j=1}^{V} \varrho_{dnl} w_{dn}^{j} \log(\Omega_{w(lj)}) + \sum_{l=1}^{D+1} \tau_l (\sum_{j=1}^{V} \Omega_{w(lj)} - 1) \tag{60}$$

Taking the derivatives with the respect to $\iota_{w(lj)}$ and setting it to zero gives 8.3.1 :

$$\Omega_{w(lj)} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j} \tag{61}$$

Beta-Liouville Parameter

The objective of this subsection is to determine estimates of the model's parameters using variational inference techniques [99].

$$\begin{aligned}
\mathcal{L}[\xi] = & \sum_{m=1}^{M} (\log(\eta(\sum_{l=1}^{D} \kappa_l)) + \log(\eta(\kappa + \iota)) - \log \eta(\kappa) - \log \eta(\iota) \\
& - \sum_{i=1}^{D} \log \eta(\kappa_i) + \sum_{i=1}^{D} \kappa_i (\Psi(\eta_{mi}) - \Psi(\sum_{l=1}^{D} \eta_{m(l)})) \\
& + \kappa(\Psi(\kappa_{m\eta}) - \Psi(\kappa_{m\eta}\iota_{m\eta})) + \iota(\Psi(\iota_{m\eta}) - \Psi(\kappa_{m\eta} + \iota_{m\eta})))
\end{aligned} \tag{62}$$

The derivative of the above equation with respect to the BL parameter is given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}[\xi]}{\partial \kappa_l} &= M(\Psi(\sum_{l=1}^{D}) - \Psi(\kappa_l)) + \sum_{m=1}^{M} (\Psi^{'}(\eta_{ml}) - \Psi(\sum_{l=1}^{D} \eta_{m(l)})) \\
\frac{\partial \mathcal{L}[\xi]}{\partial \kappa} &= M[\Psi(\kappa + \iota) - \Psi(\kappa)] + \sum_{m=1}^{M} (\Psi(\kappa_{m\eta}) - \Psi(\kappa_{m\eta} + \iota_{m\eta})) \\
\frac{\partial \mathcal{L}[\xi]}{\partial \iota} &= M[\Psi(\kappa + \iota) - \Psi(\iota)] + \sum_{m=1}^{M} (\Psi(\iota_{m\eta}) - \Psi(\kappa_{m\eta} + \iota_{m\eta}))
\end{aligned} \tag{63}$$

From the equations presented earlier, it is evident that the derivative in Eq. 54 with respect to each of the BL parameters is influenced not only by their individual values but also by their interactions with one another. Consequently, we utilize the Newton-Raphson method to address this optimization problem. To implement the Newton-Raphson method effectively, it is essential to first calculate the Hessian matrix for the parameter space, as illustrated below [48]:

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}[\xi]}{\partial \kappa_l \kappa_j} &= M(-\vartheta(i,j)\Psi^{'}(\kappa_i) + \Psi^{'}(\sum_{l=1}^{D} \kappa_l)) \\
\frac{\partial^2 \mathcal{L}[\xi]}{\partial \kappa^2} &= M(\Psi^{'}(\kappa + \iota) - \Psi^{'}(\kappa)) \\
\frac{\partial^2 \mathcal{L}[\xi]}{\partial \kappa \partial \iota} &= M\Psi^{'}(\kappa + \iota) \\
\frac{\partial^2 \mathcal{L}[\xi]}{\partial \iota^2} &= M(\Psi^{'}(\kappa + \iota) - \Psi^{'}(\iota))
\end{aligned}
\tag{64}
$$

The Hessian matrix shown above is very similar to the Hessian matrix of the Dirichlet parameters in the MPCA model and the GD parameters in GDMPCA. The matrix mentioned above can be partitioned into two distinct matrices based on the parameters κ_d, κ, and ι. The parameter derivation for each of the two halves will be identical to the Newton-Raphson model offered by MPCA and GDMPCA.

3.4.1 Inference via Collapsed Gibbs Sampling

The CGS contributes inference by utilizing a Bayesian network to estimate posterior distributions. These distributions are determined by sampling hidden variables through a process of conditional probabilities.

The CGS provides considerably faster estimation compared to the conventional Gibbs sampler, which operates in the joint space of latent variables and model parameters. The CGS operates in the latent variable space, where the model parameters θ and ϱ are marginalized out in the joint distribution $p(X, z, \theta, \varrho, w|\Omega, \Upsilon, \mu)$. The process of marginalization results in the formation of the marginal joint distribution $p(X, z, w|\Omega, \Upsilon, \mu)$, which is precisely defined as follows:

$$
p(x, z, w|\Omega, \Upsilon) = W \int_{\theta} \int_{\varphi} p(X, z, \theta, \varphi, w|\Omega, \xi) d\varphi d\theta \tag{65}
$$

Using Eq. 65, the method computes the conditional likelihoods of the hidden variables z_{ij} by taking into account the present state of all other variables, ignoring the specific variable z_{ij} itself [97]. Simultaneously, the CGS algorithm uses the conditional probability of the latent variables to assign topics to the observed words. In this context, "$-ij$" denotes counts or variables without

including z_{ij} [97]. The definition of this particular conditional probability is as follows [100]:

$$p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon) = \frac{p(z_{ij}, z^{-ij}, X, w|\Omega, \Upsilon)}{p(z^{-ij}, X, w|\Omega, \Upsilon)} \tag{66}$$

The sampling method employed in the collapsed Gibbs approach can be briefly described as an expectation problem:

$$p(z_{ij} = k|X, w, \Omega, \Upsilon) = \mathrm{E}_{p(z^{-ij}|w,X,\Omega,\Upsilon)}[p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon)] \tag{67}$$

The collapsed Gibbs sampling Beta-Liouville multinomial procedure consists of two phases for assigning documents to clusters. First, each document is assigned a random cluster for initialization. After that, each document is assigned a cluster based on the BL distribution after a set number of iterations.

The goal is to employ a network of conditional probabilities for each class to extract the latent variables from the aggregate distribution $p(X, z|w, \Omega, \Upsilon)$. The assumption of conjugacy enables the estimation of the integral in Equation 65.

$$p(X, z|w, \upsilon) = C \prod_{j=1}^{M} \left[\frac{\eta(\sum_{i=1}^{k} \kappa_i)\eta(\kappa + \iota)}{\prod_{i=1}^{k} \eta(\kappa_i)\eta(\kappa)\eta(\iota)} \right] \times \frac{\prod_{i=1}^{k} \eta(\kappa_i^{'})\eta(\kappa^{'})\eta(\iota^{'})}{\eta(\kappa^{'} + \iota^{'})\eta(\sum_{i=1}^{K} \kappa_i^{'})} \tag{68}$$

The probability of the multinomial distribution, characterized by the parameter Υ, and the probability density function of the BL distribution can be represented as:

$$\begin{aligned} p(X|\Upsilon) &= \int p(X|\theta)p(\theta|\kappa_1, \ldots, \iota, \kappa)d\theta \\ &= \int \prod_{k=1}^{k} \theta_k^{m_k} \frac{\eta(\sum_{k=1}^{k} \kappa_k)\eta(\kappa + \iota)}{\eta(\kappa)\eta(\iota)} \prod_{k=1}^{K} \frac{\theta_k^{\kappa_k - 1}}{\eta(\kappa_k)} \\ &\times (\sum_{k=1}^{K} \theta_k)^{\kappa - \sum \kappa_k} (1 - \sum_{k=1}^{K} \theta_k)^{\iota - 1} d\theta \end{aligned} \tag{69}$$

By evaluating the integral of the probability density function of the Beta-Liouville distribution

with respect to the parameter θ and adding the updated parameters obtained from the remaining integral in Eq. 71, we may represent it as a ratio of Gamma functions.

The following shows the updated parameters, where N_{jk} represents counts corresponding variables [**?**, 100]:

$$\begin{aligned} \kappa_K^{'} &= \kappa_k + \sum_{j=1}^{k} N_{jk} \\ \kappa^{'} &= \kappa + N_{jk} \\ \iota^{'} &= \iota + N_{jk} \end{aligned} \tag{70}$$

The Eq. 69 then corresponds to:

$$\begin{aligned} p(k|\kappa_1, \ldots, \kappa_k, \iota, \kappa) = \frac{\eta(\sum_{k=1}^{K} \kappa_k)\eta(\kappa + \iota)\eta(\kappa + \sum_{k=1}^{k-1} m_k)\eta(\iota + m_k)}{\eta(\kappa)\eta(\iota)\prod_{k=1}^{K} \eta(\kappa_k)\eta(\sum_{k=1}^{K}(\kappa_k + m_k))} \\ \frac{\prod_{k=1}^{K} \eta(\kappa_k + m_k)}{\eta(\kappa + \sum_{k=1}^{K-1} m_k + \iota + m_k)} \end{aligned} \tag{71}$$

The parameters $\kappa_1, \ldots, \kappa_k, \kappa$, and ι correspond to the Beta-Liouville distribution, while m_k represents the count of documents within cluster k.

After the sampling process, parameter estimation is performed. Subsequently, the empirical likelihood method [96] is utilized to validate the results using a held-out dataset. Ultimately, this process leads to the estimation of the class conditional probability $p(X|w, \Omega, \Upsilon)$ within the context of CGS:

$$p(X|w, \Omega, \Upsilon) = \prod_{ij} \sum_{k=1}^{K} \frac{1}{S} \sum_{s=1}^{S} \tilde{\theta}_{jks} \tilde{\varphi}_{kws} \tag{72}$$

The variables are then calculated as follows:

$$\tilde{\theta}_{jks} = \frac{(N_{jk} + \kappa_k)(\kappa_{jk} + \sum_{l=k+1}^{K+1} N_{jl})(N_{jk} + \iota_k)}{(a_k b_k + \sum_{l=k+1}^{K+1} N_{jl})(\kappa_j + \sum_{l=k+1}^{K+1} N_{jl})} \tag{73}$$

$$\tilde{\varphi}_{kws} = \frac{(N_{jk} + \kappa_w)(\kappa_{jw} + \sum_{l=k+1}^{K+1} N_{jl})(N_{jk} + \iota_w)}{(\kappa_w b_w + \sum_{l=k+1}^{K+1} N_{jl})(\kappa_{wj} + \sum_{l=k+1}^{K+1} N_{jl})} \tag{74}$$

where S is the size of sample.

3.5 Experimental Results

In this section, we assess the effectiveness of our proposed algorithms through two rigorous applications: topic modeling for medical texts and sentiment analysis. We evaluate each model by examining its success rate for each dataset and its perplexity [37, 10, 4, 101], a standard metric in language modeling, defined as follows:

$$prep(t_{data}) = \exp\left(\frac{-\ln p(t_{data})}{\sum_d |l_d|}\right) \tag{75}$$

where $|l_d|$ is the length of document d. A lower perplexity score indicates better generalization performance. In addition to the perplexity metric, the success rate is employed as a key performance indicator to evaluate our models, reflecting the proportion of correctly identified topics within a corpus in topic modeling. The success rate serves as a straightforward measure of a model's efficacy, capturing its ability to accurately classify documents into the correct topical categories, which is essential for effective information retrieval and knowledge discovery in the domain of text analysis. The main goal of both applications is to compare the GDMPCA, BLMPCA, and MPCA performances. The choice of these datasets is pivotal to our research as they offer a broad spectrum of analytical scenarios, from topic modeling for medical text to sentiment dataset, thus enabling a thorough investigation into the models' adaptability and accuracy. By encompassing datasets with distinct characteristics, we are able to demonstrate the strengths of our proposed models in varied contexts, highlighting their potential as a versatile tool in the field of text analysis.

3.5.1 Topic Modeling

The objective of text classification is to categorize documents into pre established subject categories, a problem extensively researched with various approaches [102, 103, 46]. Topic Modeling, a

common application in natural language processing, is used for analyzing texts from diverse sources and for document clustering [104]. It identifies key ”topics” in a text corpus using unsupervised statistical methods, where topics are keyword mixtures with a probability distribution, and documents are composed of topic mixtures [13]. The ”CMU Book Summary Dataset” was used to validate model performance, containing plot summaries and metadata for 16,559 books [105]. The models' accuracy was tested by training on various document numbers and observing the impact of latent topics on classification accuracy. Using variational Bayes inference, the models showed similar performances, but BLMPCA excelled, particularly in classifying similar classes.

In Tables 3.2, 3.3 and 3.4, we present the first three topics, the perplexity measurements and time complexity across all models compared in this study. The success rates obtained using GDMPCA, BLMPCA, and MPCA are depicted in Figure 1. These examples demonstrate that our proposed models, which incorporate Generalized Dirichlet and Beta-Liouville distributions, yield more accurate classifications in scenarios where distinct classes exhibit similarities, in contrast to the traditional MPCA which is a Dirichlet-based model. Additionally, in Tables 3.5 and 3.6, we showed the results for the collapsed Gibbs sampling.

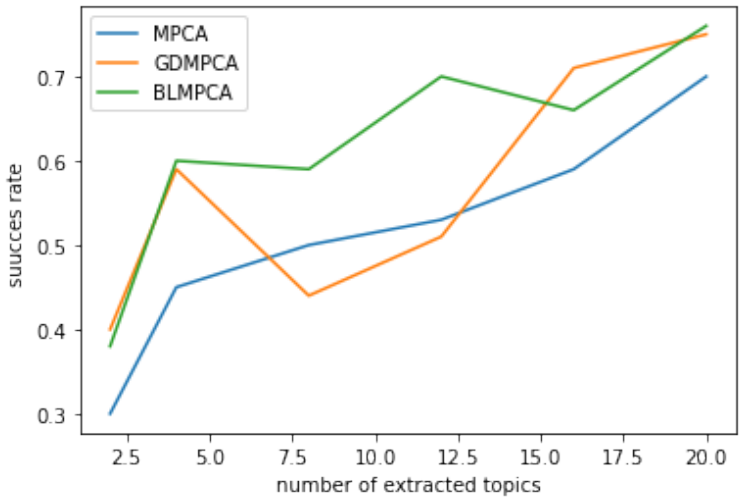


Figure 3.1: Success rate for CMU Book data

Table 3.2: Common topics identified with BLMPCA model in the CMU Book dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	girl, tells, find, two, man, when, return, after, also, finds, time, kill, later, help, killed
Topic 2	he, one, back, man, time, house, father, police, story, mother, young, school, love, time, first
Topic 3	tells, they, return, find, girl, back, one, house, story, after, dragon, find, schools, boy, jack
Topic 4	earth, world, one, human, ship, book, planet, space, human, systems, time, years, in, people, would
Topic 5	war, novel, new, world, army, story, one, group, book, states, general, british, president, first, american

Table 3.3: Comparison of the perplexity of MPCA, GDMPCA, and BLMPCA models, indicating model fit quality across different topic numbers (K) on the CMU Book dataset

K	5	10	15	20
MPCA	1455	1422	1320	1215
GDMPCA	1326	1430	1190	1178
BLMPCA	1319	1203	1198	1177

3.5.2 Topic modeling for medical text

Topic modeling plays a crucial role in navigating the complexities of health and medical text mining, despite the inherent challenges of data volume and redundancy in this domain. [106] marks a significant advancement, offering an optimized topic modeling approach that utilizes ensemble pruning. This method significantly improves the categorization of biomedical texts by enhancing precision and managing the computational challenges posed by the extensive data typical of medical documents. With vast amounts of health-related data, specialists struggle to find pertinent information, exemplified by the millions of papers on PubMed and hospital discharge records in the United States in 2015. This study utilizes the TMVAr corpus from PubMed and the TMVAr-Dataset containing health-related Twitter news to evaluate models [107, 108, 109, 110, 111, 112].

TMVAr-Dataset

The TMVar Corpus dataset, comprising 500 PubMed papers with manual annotations of various mutation mentions, is utilized to evaluate our models. Tables 3.8 and 3.9 elucidate the perplexity

Table 3.4: Time complexity comparison for MPCA, GDMPCA, and BLMPCA at varying topic levels (K) on the CMU Book dataset.

K	5	10	15	20
MPCA	107.803	140.1439	150.9242	161.7045
GDMPCA	225.04	230.544	347.056	408.064
BLMPCA	251.64	327.132	352.296	377.46

Table 3.5: Comparison of perplexity scores of MPCA, GDMPCA, and BLMPCA, reflecting model fit as topic count (K) increases on the CMU Book dataset with CGS inference.

K	5	10	15	20
MPCA	1391.5	1448.6	1516	1580
GDMPCA	1291.2	1316	1428	1413
BLMPCA	1310.4	1324.8	1416	1483.2

comparison and time complexity for the TMVAR dataset, offering insight into the performance of our proposed methods. Moreover, as shown in Table 3.7, the BLMPCA model successfully extracts pertinent topics, which are indicative of the model's nuanced analytical capabilities. Figure 3.2 further illustrates the success rate of our proposed models in comparison to the traditional MPCA, highlighting the enhanced classification accuracy achieved by our methods.

Moreover, Tables 3.10 and 3.11 present the outcomes of the collapsed Gibbs sampling. As indicated in the tables, the time complexity of this method is higher, yet the perplexity is lower.

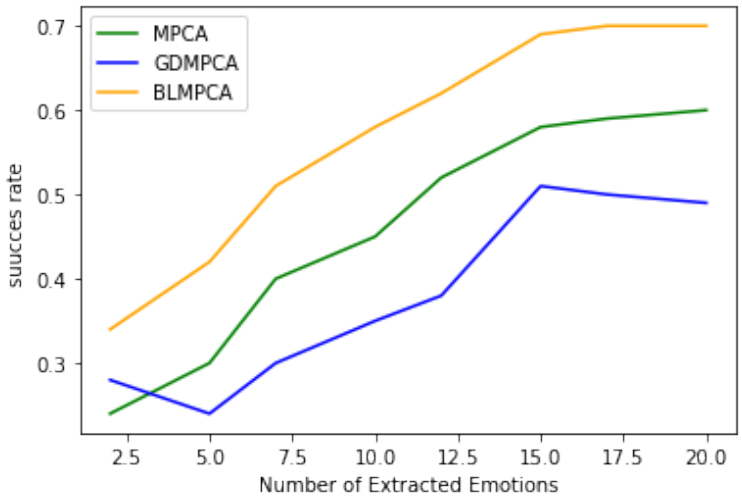


Figure 3.2: Success rate for Tmvar corpus data

Table 3.6: Time complexity comparison for MPCA, GDMPCA and BLMPCA with increasing topics (K) using CGS inference on the CMU Book dataset

K	5	10	15	20
MPCA	431.212	536.57	634.69	687.818
GDMPCA	19125.2	1138.264	2429.392	2964.51
BLMPCA	1998.84	2289.924	3018.368	3497.14

Table 3.7: Common topics identified with BLMPCA model in the TMVAR dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	mutations, mutation, gene, family, patients, iron, exon, novel, autosomal, associated
Topic 2	gene, p, cancer, polymorphism, expression, patients, associated, deletion, study, region
Topic 3	gene, patients, dna mutation, polymorphism, detected, samples, family, study, results, dna
Topic 4	dna mutation, mutations, homozygous, variants, family, ct, position, methods, associated, substitution
Topic 5	gene, patients, protein mutation, dna, exon, study, genetic , cancer, substitution, genotype

3.5.3 Sentiment Analysis

Sentiment analysis, crucial for interpreting emotions in texts from various sources, benefits from advanced methodologies beyond mere word analysis [113, 114]. Recent studies [115, 116] demonstrate the effectiveness of deep learning and text mining in capturing nuanced sentiment expressions [117]. Additionally,[118] highlights the potential of ensemble classifiers in improving sentiment classification accuracy. These innovations showcase the shift towards more complex analyses that consider semantics, context, and intensity for a more accurate sentiment understanding.

The ”Multi-Domain Sentiment Dataset” containing Amazon.com product reviews across various domains, was used for analysis [119]. This dataset, with extensive reviews on books and DVDs,

Table 3.8: Comparison of the perplexity for MPCA, GDMPCA, and BLMPCA models, indicating model fit quality across different topic numbers (K) on TMVAR dataset with variation EM inference

K	5	10	15	20
MPCA	2115	2083	1984	1977
GDMPCA	1996	1989	1968	1959
BLMPCA	1983	1965	1954	1949

Table 3.9: Time complexity comparison for MPCA, GDMPCA and BLMPCA with increasing topics (K) using variation EM inference on TMVAR dataset

K	5	10	15	20
MPCA	9.53	22.543	26.092	28.458
GDMPCA	11.83	24.843	28.392	30.758
BLMPCA	18.57	38.997	44.568	48.282

Table 3.10: Comparison of the perplexity for MPCA, GDMPCA, and BLMPCA models, indicating model fit quality across different topic numbers (K) on TMVAR dataset with CGS inference

K	5	10	15	20
MPCA	2132.5	2232.8	2376.0	2460
GDMPCA	1360.9	1182.4	1345.6	1938
BLMPCA	1938.5	11350.5	1340.5	1440

provided data for basic analysis. The applied model, using K = 8 topics, assumes each topic comprises a bag of words with specific probabilities, and each document is a mix of these topics. The model's goal is to learn the distributions of words and topics in the corpus.

We demonstrated that the overall sentiment of the dataset tends to be positive, influenced by the presence of high-frequency words with positive connotations within the corpus. This observation is substantiated by the sentiment analysis framework we employed. Tables 3.12 and 3.13 provide a detailed explanation of the perplexity measures and time complexity tested for sentiment analysis. Furthermore, the findings from the topic modeling of eight emotions and two sentiments are displayed in Tables 3.19 and 3.20. Fig. 3.3 shows that our proposed models outperform the previous model. Fig. 3.3 shows success rates for MPCA, GDMPCA, and BLMPCA on sentiment analysis, with GDMPCA and BLMPCA outperforming MPCA as the number of emotions analyzed increases. This suggests their better suitability for complex emotion detection tasks in practical applications.

Additionally, Table 3.16 and Table 3.17 display the accuracy and recall of various classifiers utilized for emotion detection. Furthermore, Tables 3.14 and 3.15 present the results for the collapsed Gibbs sampling. Table 3.18 shows the F1-scores for various classifiers, indicating the balanced harmonic mean of precision and recall for SVM, Naive Bayes, and MLP classifiers when applied with MPCA, GDMPCA, and BLMPCA models in sentiment analysis.

Table 3.11: Time complexity comparison for MPCA, GDMPCA and BLMPCA with increasing topics (K) using CGS inference on the TMVAR dataset

K	5	10	15	20
MPCA	45.74	62.89	108.20	200.63
GDMPCA	56.74	163.95	252.35	273.70
BLMPCA	165.57	336.93	376.45	392.71

Table 3.12: Comparison of the perplexity for MPCA, GDMPCA, and BLMPCA models, indicating model fit quality across different topic numbers (K) on sentiment data with variation EM inference

K	5	10	15	20
MPCA	1551	1531	1542	1529
GDMPCA	1549	1539	1524	1521
BLMPCA	1448	1540	1531	1518

Table 3.13: Time complexity comparison for MPCA, GDMPCA and BLMPCA with increasing topics (K) using variational EM inference on the sentiment analysis application

K	5	10	15	20
MPCA	130.54	169.702	182.756	195.81
GDMPCA	142.876	185.7388	200.0264	214.314
BLMPCA	158.23	205.699	221.522	237.345

Table 3.14: Comparison of the perplexity for MPCA, GDMPCA, and BLMPCA models, indicating model fit quality across different topic numbers (K) on sentiment data with CGS inference

K	5	10	15	20
MPCA	1451	1511	1589	1639
GDMPCA	1332	1393	1422	1502
BLMPCA	1316	1401	1413	1498

Table 3.15: Time complexity comparison for MPCA, GDMPCA and BLMPCA with increasing topics (K) using CGS inference on the sentiment analysis application

K	5	10	15	20
MPCA	830.54	1069.702	1282.756	1495.81
GDMPCA	924.451	1258.78	1319.46	1383.17
BLMPCA	1085.42	1264.24	1390.12	1473.623

Table 3.16: Accuracy comparisons for sentiment analysis classifiers

Classifier	SVM	NaiveBayes	MLP
MPCA	0.62	0.68	0.67
GDMPCA	0.80	0.85	0.87
BLMPCA	0.83	0.88	0.88

Table 3.17: Recall metrics for SVM, Naive Bayes, and MLP classifiers using MPCA,GDMPCA and BLMPCA in sentiment analysis.

Classifier	SVM	NaiveBayes	MLP
MPCA	0.61	0.59	0.66
GDMPCA	0.79	0.76	0.85
BLMPCA	0.85	0.82	0.89

Table 3.18: F1-score metrics for SVM, Naive Bayes, and MLP classifiers using MPCA, GDMPCA, and BLMPCA in sentiment analysis.

Classifier	SVM	Naive Bayes	MLP
MPCA	0.6195	0.6041	0.6697
GDMPCA	0.7999	0.7701	0.8593
BLMPCA	0.8593	0.8313	0.8999

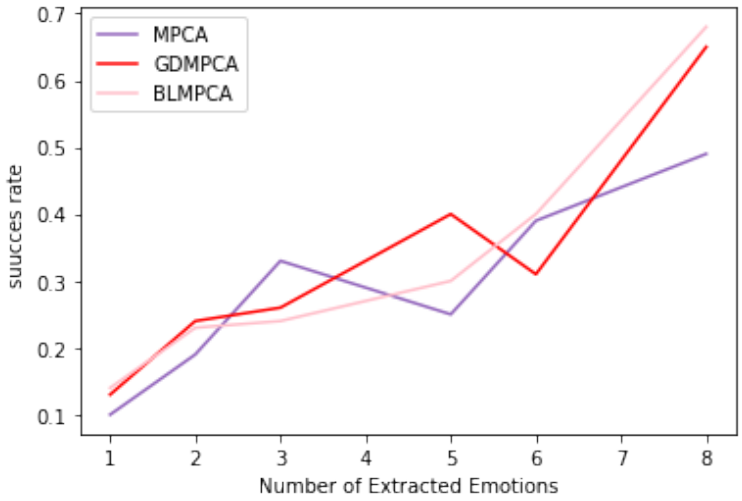


Figure 3.3: Success rate for Sentiment Dataset

Table 3.19: Frequency of emotions identified in text data via topic modeling

Emotions	Count
satisfied	78901
angry	21345
happy	6521
joy	82345
disgust	7125
Perfect	45459
Tearful	3451
sad	4387

Table 3.20: The counts of positive, negative, and unlabeled sentiments identified through sentiment analysis.

Sentiment	Count
Positive	213232
Negative	36308
Unlabeled	23451

Chapter 4

Bi-Directional Recurrent Attentional Topic Model Using Flexible Priors

4.1 Introduction

Advances in information technology combined with social media, where people now share knowledge and skills, have created an information revolution. Daily created websites are becoming great resources for data science and analytics in addition to being stores of important information generally hidden in documents. One crucial task in machine learning is modeling documents into a vector space. Therefore, it is essential to carefully analyze how to use document components, including words, phrases, and paragraphs. Due to the complexity and variability of massive data collections, processing such unstructured records necessitates efficient machine learning techniques. In topic modeling, such collections are summarized as documents that use the bag-of-words method [120] to perform on count data. The goal is to efficiently generate topics to make accurate predictions on unseen documents in tasks like retrieval and classification.

Capturing and accurately reflecting the complexity of textual data is a significant difficulty in statistical analysis. To address this issue, scientific document analysis and mining have used approaches like data projection [3, 4] and clustering. Recently, machine learning and deep learning have been used for text classification [5], summarization [6], segmentation [7], topic modeling [8], and sentiment analysis [9]. Probabilistic topic models, such as PLSA [11, 12] and LDA [13], offer

effective document representation by analyzing word distributions within document corpora, with LDA improving on PLSA by incorporating Dirichlet distribution variables [121].

In summary, using the bag-of-words assumption, documents comprise collections of words with varying frequencies. Therefore, much effort has concentrated on obtaining improved representations for documents in a variety of real-world applications. When confronted with a collection of documents, probabilistic topic models give a suite of algorithms for obtaining appropriate representations. However, when the document set is noisy, topic models can lead to errors in document representation [37]. Additionally, topic models commonly assume that words are interchangeable, which facilitates fast inference on extensive collections of text [37]. A document is composed of a hierarchical arrangement of words, sentences, and paragraphs. Some methods, including n-gram language modeling [81] and recurrent neural networks (RNN) [56] for language modeling, have depicted a document as a sequence of words [82, 55].

It is significant to mention that sentence sequences can be found in the documents; therefore, if the sequential information between sentences is accessible, simply evaluating word information is insufficient [37]. As a result, a document might be perceived as a forward and backward-moving sentence sequence. Besides coherence among sentences on a particular topic, it is reasonable to assume that a sentence has different degrees of association with its neighboring sentences. The integration of sequential nature and weight signals of sentences in a text for topic coherence at the sentence level is a vital and essential aspect of the topic model's architecture [37]. Recurrent neural networks (RNNs) and their derivatives, such as long short-term memory (LSTM) [122], have achieved significant success in document modelling by utilizing sequence information [123], in which texts are seen as word sequences. Furthermore, deep neural networks have used the attentional process to obtain data from the relevant parts of text data [124, 37].

Techniques like LDA [13] are utilized to create models for unstructured texts, which are based on the concept that the words found in a document are a result of a combination of hidden topics. Each topic is represented by a distribution of the vocabulary used in the text. The authors in [37] proposed the "Bi-Directional Recurrent Attentional Bayesian Process (bi-RABP)" [37] model to eliminate the discrepancy in sequential data handling and authorize local recurrent information

transmission across a sequence. The bi-RATM can model sentence sequences by taking into account sentence dependencies as well as attention signals from two successive sentences [37]. The common issue of these models is using the Dirichlet prior in their framework. To overcome the constraints of the Dirichlet prior, more flexible priors like GD and BL can be utilized as alternatives to reconstruct the generative process. It is noteworthy that the topic components are independent under the Dirichlet distribution, which removes topic correlation from the model. Since it does not allow any dependency between distinct topics, the LDA was unable to provide a natural method of arranging documents [49]. This structure improves the smoothness of grouping and compression procedures. While conjugate priors have been employed in generating closed-form posteriors due to their simple structure, some topic modelling methodologies, such as the correlated topic model [49], have advocated the use of non-conjugate priors as alternatives [125, 126].

This chapter discusses the proposed models, which include the GD-bi-RATM and BL-bi-RATM. We first describe the properties of the fitting distribution for each proposed model, then estimate the parameters of each distribution using variational inference, and finally provide the complete learning algorithm.

4.2 Generalized Dirichlet Bi-Directional Topic Modeling

According to [91], when clustering count data using mixture models, the generalized Dirichlet (GD) distribution can be a suitable alternative to the Dirichlet distribution. This is because the GD distribution is also a conjugate prior to the multinomial distribution, and it has a more general covariance matrix [127]. To extend the bi-RATM model based on the GD assumption, the variational Bayes method will be employed. It is expected that the GD-bi-RATM will perform better than the Dirichlet-based model because the Dirichlet distribution is a special case of the more general GD distribution, as stated in [92].

The preceding phrases have a forward influence on the sentence contents, which matches with the RABP's motivation. The future sentences, which can be represented by another RABP, have an impact on the sentence contents when read backward. Thus, the method used to derive the topic distribution of a sentence involves analyzing the sentences that precede and follow it. Similar to the

other models, a basic bag-of-words language model is used for each sentence[13, 37].

In our analysis, a corpus refers to a group of M documents represented as $D = \{w_1, w_2, \ldots, w_M\}$. Each individual document w_m is composed of a sequence of N_m words, which can be denoted as $w_m = (w_{m1}, \ldots, w_{mN_m})$.

θ^{sj} signify the topic distribution over K latent topics of the j^{th} sentence $\boldsymbol{s}_j$ in document D.

The binary vector w_n is selected from a vocabulary that contains V words [46]. In order to generate each word in the document using the GD-bi-RATM model, the following steps are taken, where c is a binary vector with $d+1$ dimensions representing topics.

$$\begin{aligned} m &\sim GD(\xi) \\ \nu &\sim GD(\zeta) \\ z &\sim Multinomial(m, L) \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{76}$$

If the i^{th} topic is chosen, $z_i^n = 1$; in other cases, $z_i^n = 0$. $m = (m_1, \ldots, m_{d+1})$, where $m_{d+1} = 1 - \sum_{i=1}^{d} m_i$.

The multinomial probability $p(w_n|z_n, \Omega_w)$ is conditional on z_n. $GD(\xi)$, and $GD(\zeta)$ is a d-variate GD distribution with parameters $\xi = (a_1, b_1, \ldots, a_d, b_d)$, and $(\zeta) = (aa_1, bb_1, \ldots, aa_d, bb_d)$ and probability distribution function p, where $\gamma_i = b_i - a_{i+1} - b_{i+1}$ [46]:

$$p(m_1, \ldots, m_d|\xi) = \prod_{i=1}^{d} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} m_i^{a_i - 1}(1 - \sum_{j=1}^{i} m_j)^{\gamma_i} \tag{77}$$

When b_i is equal to $a_{(i+1)} + b_{(i+1)}$, the GD distribution can be converted into the Dirichlet distribution. The GD distribution has a mean and variance as shown below [92]:

$$E(m_i) = \frac{a_i}{a_i + b_i} \prod_{k=1}^{i-1} \frac{b_k}{a_k + b_k} \tag{78}$$

$$var(m_i) = E(m_i)\left(\frac{a_i+1}{a_i+b_i+1}\prod_{k=1}^{i-1}\frac{b_k+1}{a_k+b_k}+1-E(\theta_i)\right) \tag{79}$$

The covariance between m_i and m_j can be expressed as follows

$$cov(m_i, m_j) = E(m_j)\left(\frac{a_i}{a_i+b_i+1}\prod_{k=1}^{i-1}\frac{b_k+1}{a_k+b_k}+1-E(m_i)\right) \tag{80}$$

The GD distribution's covariance matrix is more flexible than the Dirichlet distribution's; accordingly, compared to the Dirichlet distribution, the generalized Dirichlet distribution is characterized by more significant variability. Additionally, it includes an extra set of parameters that provide $d-1$ more degrees of freedom, making it more suitable for real-world data. Indeed, the GD distribution fits count data better than the commonly used Dirichlet distribution [93].

4.2.1 Model Inference

Inference in the GD-bi-RATM model involves a significant challenge in accurately estimating the posterior distribution of latent variables based on the observed data. For instance, consider a sentence s_j within a document d consisting of a set of words N_j. In the proposed method, calculating the posterior distribution of latent variables can be achieved through the following steps:

$$P(\epsilon^j, z|s_j, \mu, \pi, \zeta) = \frac{P(\epsilon^j, z, s_j|\mu, \pi, \zeta)}{P(s_j|\mu, \pi, \zeta)} \tag{81}$$

Based on the conjugate relationship between the prior over-the-topic distribution and the topic assignment distribution, Gibbs sampling methods can be utilized to train common topic models. However, bi-RATM makes the posterior distribution computationally intractable because the prior topic distribution of a phrase is incompatible with this condition. Consequently, variational inference is utilized to provide an approximation of the posterior of a sentence to lower the Kullback–Leibler (KL) divergence between the variational and real posterior distributions [94, 128]. The variational technique reinterprets the inference problem as an optimization challenge to come up with an approximation of the posterior distribution [21]. According to the author in [66], it is possible to increase the evidence lower bound (ELBO) [129, 37] by increasing the variational posterior

probability and decreasing the KL divergence between the variational probability and the true posterior probability. In variational inference, the posterior distribution is approximated using a set of variational distributions with free variational parameters. The purpose is to bring various variational distributions as close to the true posterior as possible. However, variational inference is employed to approximate the posterior distribution of the sentence and minimize the Kullback-Leibler (KL) divergence between the true posterior distribution and the approximate posterior distribution. To estimate the posterior distribution of each sentence $\boldsymbol{s}_j$ containing N_j words in document d, the fully factorized variational distribution is applied in the following [37]:

$$q(\phi, \epsilon, z|\Omega, \xi, \gamma) = q(\phi|\Omega) \prod_{j=1}^{s_j} q(\epsilon|\xi) \prod_{n=1}^{N_j} q(z_n|\gamma_n) \tag{82}$$

Instead of using the posterior distribution $p(\epsilon^j, z, s_j|\mu, \pi, A)$, we employ an optimization method to determine the variational parameters γ and Φ. The process of optimization will be explained in detail in the following. Jensen's inequality is used to bound the log-likelihood and eliminate the parameters γ and Φ for simplicity [94]:

$$\begin{aligned}
\log p(s_j|\xi, \Omega) &= \log \int \sum_z p(\epsilon, z, s_j|\xi, \Omega) dm \\
&= \log \int \sum_z \frac{p(\epsilon, z, s_j|\xi, \Omega) q(\epsilon, z)}{q(\epsilon, z)} d\epsilon \\
&\geq \int \sum_z \log p(\epsilon, z, s_j|\xi, \Omega) q(\epsilon, z) dm \\
&- \int \sum_z q(\epsilon, z) \log q(\epsilon, z) dm \\
&= \mathbb{E}[\log p(\epsilon, z, s_j|\xi, \Omega)] - \mathbb{E}[\log q(\epsilon, z)]
\end{aligned} \tag{83}$$

Thus, the lower bound on the log-likelihood for any variational distribution $q(\epsilon, z|\gamma, \Phi)$ can be obtained using Jensen's inequality.

To measure the difference between the variational and actual posterior probabilities, the KL divergence is calculated as the disparity between the expressions on the left and right-hand sides of equation 83. If we denote $\mathcal{L}(\mu, \pi, \phi; \xi, \gamma, \Omega)$ the right side of Eq. 83, to incorporate the variational parameters dependency, we obtain:

$$
\begin{aligned}
\mathcal{L}^d(\mu,\pi,\phi;\xi,\gamma,\Omega) &= \sum_{j=1}^{s_j} \mathcal{L}^{s_j}(\mu,\pi;\xi,\gamma) + \mathbb{E}_q[\log p(\phi|\zeta)] - \mathbb{E}_q[\log q(\Omega)] \\
&= \sum_{j=1}^{s_j} \mathbb{E}_q[\log p(\epsilon|\pi)] + \sum_{j=1}^{s_j}\sum_{n=1}^{N_j} \mathbb{E}_q[\log p(z|\epsilon,\theta^{2C+1})] \\
&+ \sum_{j=1}^{s_j}\sum_{n=1}^{N_j} \mathbb{E}_q[\log p(w_n|z_n,\mu)] \\
&- \sum_{j=1}^{s_j} \mathbb{E}_q[\log q(\epsilon)] - s_j \cdot \mathbb{E}_q[\log q(z)] + \mathbb{E}_q[\log p(\phi)]
\end{aligned} \tag{84}
$$

The first term in the equation can be broken down into the log probability of sentence s_j:

$$
\begin{aligned}
\mathcal{L}^{s_j}(\mu,\pi;\xi,\gamma) &= \mathbb{E}_q[\log p(\epsilon|\pi)] - \mathbb{E}_q[\log q(z)] + \sum_{n=1}^{N_j} \mathbb{E}_q[\log p(z_n|\epsilon,\theta^{j-c:j-1},\theta^{j+1:j+c},\phi)] \\
&+ \sum_{n=1}^{N_j} \mathbb{E}_q[\log p(w_n|z_n,\mu)] - \mathbb{E}_q[\log q(\epsilon)] \\
&= \sum_{n=1}^{N_j}\sum_{k=1}^{K} \gamma_{nk} \sum_{l=1}^{2c+1} \log \theta_{lk}^{2c+1} \frac{\xi_l}{\sum_{l'=1}^{2c+1}\xi_{l'}} + \log\Gamma\Big(\sum_{l=1}^{2c+1}\pi_l\Big) \\
&+ \sum_{l=1}^{2c+1}(\pi_l - 1)\Big(\Psi(\epsilon_l) - \Psi\Big(\sum_{l'=1}^{2c+1}\xi_{l'}\Big)\Big) - \sum_{l'=1}^{2c+1}\log\Gamma(\pi_{l'}) \\
&+ \sum_{n=1}^{N_j}\sum_{k=1}^{K}\sum_{v=1}^{V} \gamma_{nk} w_n^v \log\mu_{kv} - \sum_{n=1}^{N_j}\sum_{k=1}^{K}\gamma_{nk}\log\gamma_{nk} - \log\Gamma\Big(\sum_{l=1}^{2c+1}\xi_l\Big) \\
&+ \sum_{l=1}^{2c+1}\log\Gamma(\xi_l) - \sum_{l=1}^{2c+1}(\xi_l - 1)\Big(\Psi(\xi_l) - \Psi\Big(\sum_{l'=1}^{2c+1}\xi_{l'}\Big)\Big)
\end{aligned} \tag{85}
$$

where V is the size of dictionary.

It should be noted that while computing $\mathbb{E}_q[\log p(z_n|\epsilon,\theta^{2c+1})_k]$ is challenging, we can obtain

its lower bound for sentence s_j as follows:

$$\begin{aligned}\sum_{n=1}^{N_j} \mathbb{E}_q[\log p(z_n|\epsilon, \theta^{2c+1})] &= \sum_{n=1}^{N_j}\sum_{k=1}^{K} \gamma_{nk}\mathbb{E}_q[\log(\epsilon \times \theta^{2c+1})_k] \\ &\geq \sum_{n=1}^{N_j}\sum_{k=1}^{K}\sum_{l=1}^{2c+1} \log \theta_{lk}^{2c+1}\mathbb{E}_q[\epsilon] \\ &= \sum_{n=1}^{N_j}\sum_{k=1}^{K} \gamma_{nk}\sum_{l=1}^{2c+1} \log \theta_{lk}^{2c+1}\frac{\xi_l}{\sum_{l'}\xi_{l'}}\end{aligned} \tag{86}$$

In order to simplify the inference process, we introduce an augmented matrix notation for the term $\mathbb{E}_q[\log p(z_n|\epsilon, \theta^{j-c:j-1}, \theta^{j+1:j+c}, \phi)]$. Here, $\theta^{j-c:j-1}$ and $\theta^{j+1:j+c}$ refer to the topic matrices of the preceding and succeeding sentences for the sentence s_j:

$$\sum_{n=1}^{N_j} \mathbb{E}_q[\log p(z_n|\epsilon, \theta^{j-c:j-1}, \theta^{j+1:j+c}, \phi)] = \mathbb{E}_q[\log p(z_n|\epsilon, (\theta^{j-c:j-1}/\theta^{j+1:j+c}/\phi))] \tag{87}$$

Finally, by plugging in Eqs. 86 and 84, $\mathcal{L}^d(\mu, \pi, \zeta; \xi, \gamma, \Omega)$ can be computed as:

$$
\begin{aligned}
\mathcal{L}^d(\mu, \pi, \zeta; \xi, \gamma, \Omega) = s_i \cdot \log\Gamma\Big(\sum_{l=1}^{2c+1} \pi_l\Big) - \\
s_i \cdot \sum_{l'=1}^{2c+1} \log\Gamma(\pi_{l'}) + s_i \cdot \sum_{l=1}^{2c+1} (\pi_l - 1)\Big(\Psi(\xi_l) - \Psi\Big(\sum_{l'=1}^{2c+1} \xi_{l'}\Big)\Big) \\
+ \sum_{j=1}^{s_j} \sum_{n=1}^{N_j} \sum_{k=1}^{K} \gamma_{nk} \sum_{l=1}^{2c+1} \log \theta_{lk}^{2c+1} \frac{\xi_l}{\sum_{l'=1}^{2c+1} \xi_{l'}} \\
+ \sum_{j=1}^{s_j} \sum_{n=1}^{N_j} \sum_{k=1}^{K} \sum_{v=1}^{V} \gamma_{nk} w_n^v \log \mu_{kv} \\
- s_i \cdot \log\Gamma\Big(\sum_{l=1}^{2c+1} \xi_l\Big) + s_i \cdot \sum_{l=1}^{2c+1} \log(\xi_l) \\
- s_i \cdot \sum_{l=1}^{2c+1} (\xi_l - 1)\Big(\Psi(\xi_l) - \Psi\Big(\sum_{l'=1}^{2c+1} \xi_{l'}\Big)\Big) \\
- \sum_{j=1}^{s_j} \sum_{n=1}^{N_j} \sum_{k=1}^{K} \gamma_{nk} \log \gamma_{nk} + \log\Gamma\Big(\sum_{k=1}^{K} \zeta_k\Big) - \sum_{k=1}^{K} \log\Gamma(\zeta_k) \\
+ \sum_{k=1}^{K} (\zeta_K - 1)\Big(\Psi(\Omega) - \Psi\Big(\sum_{k=1}^{K} \Omega\Big)\Big) \\
- \log\Gamma\Big(\sum_{k=1}^{K} \Omega_k\Big) + \sum_{k=1}^{K} \log\Gamma(\Omega_k) \\
- \sum_{k=1}^{K} (\Omega_k - 1)\Big(\Psi(\Omega) - \Psi\Big(\sum_{k=1}^{K} \Omega\Big)\Big)
\end{aligned}
\tag{88}
$$

The attention signal is connected to its corresponding sentence. Therefore, for each sentence, we aim to maximize ξ as shown in Equation 88. Based on Eq. 88, the equivalent objective function for phrase s_j is as follows:

$$
\begin{aligned}
\mathcal{L}_{[\xi]}^{s_j} = \sum_{l=1}^{2c+1} (\pi_l - 1)\Big(\Psi(\xi_l) - \Psi\Big(\sum_{l'=1}^{2c+1} \xi_{l'}\Big)\Big) - \log\Gamma\Big(\sum_{l=1}^{2c+1} \xi_l\Big) \\
+ \sum_{l=1}^{2c+1} \log\Gamma(\xi_l) - \sum_{l=1}^{2c+1} (\xi_l - 1)\Big(\Psi(\xi_l) - \Psi\Big(\sum_{l'=1}^{2c+1} \xi_{l'}\Big)\Big) \\
+ \sum_{n=1}^{N_j} \sum_{k=1}^{K} \gamma_{nk} \sum_{l=1}^{2c+1} \log \theta_{lk}^{2c+1} \frac{\xi_l}{\sum_{l'=1}^{2c+1} \xi_l'}
\end{aligned}
\tag{89}
$$

We will ignore s_j, and then we compute the derivative of ξ as:

$$\begin{aligned}\mathcal{L}^{'}(\xi_l) = \Psi^{'}(\xi_l)\Big(\sum_{i=1}^{2c+1}\pi_i - \xi_l\Big) + \\ \sum_{n=1}^{N_j}\sum_{k=1}^{K}\gamma_{nk}\Big(\frac{\log\theta_{lk}^{2c+1}(\sum_{j=1}^{2c+1}\xi_j) - \sum_{i=1}^{2c+1}\log\theta_{ik}^{2c+1}\xi_i}{\sum_{j'=1}^{2c+1}\xi_{j'}^2}\Big) \\ -\Psi^{'}\Big(\sum_{i=1}^{2c+1}\xi_i\Big)\sum_{l=1}^{2c+1}\Big(\sum_{i=1}^{2c+1}\pi_i - \xi_l\Big)\end{aligned} \tag{90}$$

Our goal is to optimize Eq. 88 concerning γ. To achieve this, we define the objective function for γ as:

$$\begin{aligned}\mathcal{L}^{'}(\xi_l) = \sum_{n=1}^{N_j}\sum_{k=1}^{K}\gamma_{nk}\sum_{l=1}^{2c+1}\log\theta_{lk}^{2c+1}\frac{\xi_l}{\sum_{l'=1}^{2c+1}\xi_{l'}} - \sum_{n=1}^{N_j}\sum_{k=1}^{K}\gamma_{nk}\log\gamma_{nk} \\ + \sum_{n=1}^{N_j}\sum_{k=1}^{K}\sum_{v=1}^{V}\gamma_{nk}w_n^v\log\mu_{kv}\end{aligned} \tag{91}$$

To optimize ξ with respect to γ, the objective function is constructed as shown in Eq. 88. The digamma function Ψ, which is the logarithmic derivative of the Gamma function, is used in the function. The gradient descent method is utilized to compute the estimate of ξ. The topic distribution θ^j for the sentence $\boldsymbol{s}_j$ can be updated once the attention signals are learned for that sentence:

$$\theta_k^j = \sum_{l=1}^{2c+1}\frac{\xi_l}{\sum_{l'}\xi_l'}\theta_{lk}^{2c+1} \tag{92}$$

Variational Update for Word Assignment

By applying variational inference to optimize the lower bounds on Equation 89, we obtain the subsequent updating equations for the variational multinomial. In order to find ϕ_{nl}, we proceed to maximize with respect to ϕ_{nl} so we have following equations [46]:

$$L[\phi_{nl}] = \phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \phi_{nl} \log \beta_{w(lv)} - \phi_{nl} \log \phi_{nl} + \lambda_n(\sum_{ll=1}^{d+1} \phi_{n(ll)} - 1) \tag{93}$$

and

$$L[\phi_{n(d+1)}] = \phi_{n(d+1)}(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d) + \phi_{n(D+1)} \log \beta_{(d+1)v} - \phi_{n(d+1)} \log \phi_{n(d+1)} + \lambda_n(\sum_{ll=1}^{d+1} \phi_{n(ll)} - 1) \tag{94}$$

and therefore we have:

$$\frac{\partial L}{\partial \phi_{nl}} = (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \log \beta_{lv} - \log \phi_{nl} - 1 + \lambda_n \tag{95}$$

and

$$\frac{\partial L}{\partial \phi_{n(d+1)}} = (\Psi(\gamma_d) - \Psi(\gamma_d + \delta_d)) + \log \beta_{(d+1)v} - \log \phi_{n(d+1)} - 1 + \lambda_n \tag{96}$$

Setting the above equation to zero leads to

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} \tag{97}$$

$$\phi_{n(d+1)} = \beta_{(d+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d))} \tag{98}$$

Considering that $\sum_{ll=1}^{d+1} \phi_{n(ll)} = 1$ for the normalization factor, we have:

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} \beta_{lv} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} + \beta_{(d+1)v} e^{(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d))}} \tag{99}$$

For each word w_n in sentence s_j, a topic index z_n is assigned, and γ_l and δ_l represent the variational parameters associated with the likelihood that the word w_n is assigned to topic k.

The updated equations are as follows:

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \frac{\xi_l}{\sum_l \xi_l} \tag{100}$$

$$\delta_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \frac{\xi_l}{\sum_l \xi_l} \tag{101}$$

Variational Update for Document Embedding

The focus of bi-RATM as a topic model is to extract each document's topic distribution for document embedding. $G_0 = \delta_\zeta$ is defined in the preceding variational inference and update the topic distribution for the document, ζ, which is the embedding of one document, is updated. As mentioned in [37], ϵ_{2c+1} is considered as the attention signal for ν_d. The maximization of $\mathcal{L}^d(\mu, \pi, \zeta; \xi, \gamma, \Omega)$ with respect to ζ^d and Ω is performed when μ, π, ξ, γ are fixed in the alternating optimization. The following equation is derived by setting the derivative with respect to Ω to 0, where α_l and β_l are initialized by GD [37].

$$\Lambda = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \frac{\xi_l}{\sum_l \xi_l} \tag{102}$$

$$\Theta = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \frac{\xi_l}{\sum_l \xi_l} \tag{103}$$

According to Eqs. 102 and 103, ν may be obtained as a normalized Ω:

$$\nu = \frac{\Gamma(\Lambda)\Gamma(\Theta)}{\Gamma(\Lambda + \Theta)} \tag{104}$$

Parameter Estimation

The terms of Eq. 84 containing the GD parameters ξ are chosen:

$$\begin{aligned} \mathcal{L}[\xi] = &\sum_{m=1}^{M} (\log(\Gamma(\alpha_l + \beta_l)) - \log\Gamma(\alpha_l)) - \log(\Gamma(\beta_l))) \\ &+ \sum_{m=1}^{M} (\alpha_l(\Psi(\gamma_{ml} - \Psi(\gamma_{ml} + \delta_{ml})) + \beta_l(\Psi(\delta_{ml}) - \Psi(\delta_{ml} - \gamma_{ml}))) \end{aligned} \tag{105}$$

The above equation's derivative with respect to GD parameters yields:

$$\frac{\partial \mathcal{L}[\xi]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M}(\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{106}$$

and

$$\frac{\partial \mathcal{L}[\xi]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M}(\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{107}$$

In order to solve the Newton-Raphson equation, it is necessary to obtain the Hessian matrix in the parameter space, which can be used in the optimization process:

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \alpha_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\alpha_l)] \tag{108}$$

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \beta_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\beta_l)] \tag{109}$$

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \alpha_l \beta_l} = M[\Psi^{'}(\alpha_l + \beta_l)] \tag{110}$$

The update equation of μ is:

$$\mu = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{d_{nl}} w_{d_n}^{j} \tag{111}$$

4.3 Beta-Liouville Bi-Directional Topic Modeling

For the Beta-Liouville Bi-Directional model (BL-bi-RATM), we define the same scenario as a collection of documents with the same assumption described in the GD-bi-RATM section. Hence, we have the following procedure for the model on every single word of the document. The BL-bi-RATM model proceeds with generating every single word, given a sentence s_j from the document d, with the following steps, where the vector c is a binary vector of topics with $(d+1)$ dimensions, and it is defined as follows:

$$\begin{aligned} \tau &\sim BL(\Upsilon) \\ \Omega_k &\sim BL(\iota) \\ z &\sim Multinomial(\tau, L) \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{112}$$

If the i^{th} topic is chosen, $z_i^n = 1$; in other cases, $z_i^n = 0$. z_n is a $(D+1)$-dimensional binary of topics. τ is defined as $\tau = (\tau_1, \tau_2, ..., \tau_{D+1})$ and $\tau_{D+1} = 1 - \sum_{i=1}^{D} \tau_i$.

For a selected topic, a multinomial prior w over the vocabulary of words is chosen such that $\Omega_{w_{ij}} = p(w^j = 1 | z^i = 1)$, from which each word is randomly selected. The probability $p(w_n | z_n, \Omega_w)$ is a multinomial probability based on z_n, and $BL(\Upsilon)$ is a d-variate Beta-Liouville distribution that has parameters $\Upsilon = (\alpha_1, ..., \alpha_D, \alpha, \beta)$ and a probability distribution function that can be expressed as:

$$\begin{aligned} P(\theta_1, \ldots, \theta_D | \Upsilon) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{\theta_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times \\ \left(\sum_{d=1}^{D} \theta_d \right)^{\alpha - \sum_{l=1}^{D} \alpha_l} \times \left(1 - \sum_{l=1}^{D} \theta_l \right)^{\beta - 1} \end{aligned} \tag{113}$$

The Dirichlet distribution is the special case of BL if $\beta_d = \alpha_{d+1} + \beta_{d+1}$ [**?**, 46].

To describe the BL distribution, the following statistical properties are used: mean, variance, and covariance.

$$E(\theta_d) = \frac{\alpha}{\alpha+\beta} \frac{\alpha_d}{\sum_{d=1}^{D} \alpha_d} \tag{114}$$

$$\begin{aligned} var(\theta_d) = \left(\frac{\alpha}{\alpha+\beta}\right)^2 \frac{\alpha_d(\alpha_d+1)}{(\sum_{m=1}^{D} \alpha_m)(\sum_{m=1}^{D} \alpha_m + 1)} \\ - E(\theta_d)^2 \frac{\alpha_d^2}{(\sum_{m=1}^{D} \alpha_m)^2} \end{aligned} \tag{115}$$

and the covariance between θ_l and θ_k is given by:

$$Cov(\theta_l, \theta_k) = \frac{\alpha_l \alpha_k}{\sum_{d=1}^{D} \alpha_d} \left(\frac{\frac{(\alpha+1)(\alpha)}{(\alpha+\beta+1)(\alpha+\beta)}}{\sum_{d=1}^{D} \alpha_d + 1} - \frac{\frac{\alpha}{\alpha+\beta}}{\sum_{d=1}^{D} \alpha_d} \right) \tag{116}$$

The preceding equation demonstrates that the covariance matrix of the Beta-Liouville distribution is more general compared to the covariance matrix of the Dirichlet distribution.

4.3.1 Model Inference

For the parameter estimation of BL-bi-RATM, first the parameter Ω was estimated by the Beta-Liouville prior on τ using parameters Υ [48]. The likelihood model for the BL-bi-RATM is given as follows:

$$\begin{aligned} p(\tau, w|\Upsilon, \Omega) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)} z^{L}_{w_{1,1}, w_{1,2} \ldots, w_{k,1}, w_{1,J} \ldots w_{K,J}} \big[\frac{1}{\Gamma(\alpha_d)} \tau_k^{\alpha_d - 1} + \\ \sum_k \tau_k^{\alpha - \sum_d \alpha_d} + (1 - \sum_k \tau_k)^{\beta - 1} \big] \prod_{k,j} \tau_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \end{aligned} \tag{117}$$

For the Beta-Liouville priors, we have the following:

$$\begin{aligned} \tau &\sim BL(\Upsilon) \\ \Omega_k &\sim BL(\iota) \end{aligned} \tag{118}$$

In the following step, we will estimate the parameters for Ω using the Beta-Liouville prior and the Hessian matrix.

As we explained in Section 4.2, we should estimate the model parameters (Υ, Ω) and the variational parameters (γ, Φ), according to Eqs. 84 and 83, to find τ_{nl}. We then proceed to maximize

with the respect to τ_{nl} so we have following equations:

$$\begin{aligned}
\mathcal{L}(\gamma, \Phi; \Upsilon, \Omega) &= \log(\Gamma(\sum_{d=1}^{D} \alpha_d)) + \log(\Gamma(\alpha + \beta)) - \log(\Gamma(\alpha)) \\
&\quad - \log(\Gamma(\beta)) - \sum_{d=1}^{D} \log \Gamma(\alpha_d) + \sum_{d=1}^{D} \alpha_d (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l) \\
&\quad + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(\Psi(\beta_\gamma) \\
&\quad - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
&\quad + \sum_{n=1}^{N} \sum_{d=1}^{D} \tau_{nd} (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
&\quad + \sum_{n=1}^{N} \tau_{n(D+1)} (\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
&\quad + \sum_{n=1}^{N} \sum_{l=1}^{D+1} \sum_{j=1}^{V} \tau_{nl} w_n^j \log(\Omega_{lj}) \\
&\quad - \big(\log(\Gamma(\sum_{l=1}^{D} \alpha_l)) + \log(\Gamma(\alpha + \beta)) - \log \Gamma(\alpha) - \log \Gamma(\beta) \\
&\quad - \sum_{i=1}^{D} \log \Gamma(\alpha_i) \\
&\quad + \sum_{i=1}^{D} \alpha_i (\Psi(\gamma_{mi}) - \Psi(\sum_{l=1}^{D} \gamma_{\tau(l)})) + \alpha_(\Psi(\alpha_{\tau\gamma}) \\
&\quad - \Psi(\alpha_{\tau\gamma} \beta_{\tau\gamma})) + \beta(\Psi(\beta_{\tau\gamma}) - \Psi(\alpha_{\tau\gamma} + \beta_{\tau\gamma}))\big) \\
&\quad - \big(\sum_{n=1}^{N} \sum_{l=1}^{D+1} \tau_{nl} \log(\tau_{nl}) \big)
\end{aligned} \tag{119}$$

To find τ_{nl}, we proceed to maximize with respect to ϕ_{nl}:

$$\begin{aligned}
\mathcal{L}[\tau_{nl}] &= \tau_{nl}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \tau_{nl} \log \beta_{w(iv)} - \tau_{nl} \log(\tau_{nl}) \\
&\quad + \lambda_n (\sum_{l=1}^{D} \tau_{nl} - 1)
\end{aligned} \tag{120}$$

Therefore we have:

$$\frac{\partial \mathcal{L}}{\partial \phi_{nl}} = (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \log \beta_{w(iv)} - \log \phi_{nl} - 1 + \lambda_n \tag{121}$$

The next step is to optimize Eq. 119 to find the updated equations for the variational; we separate the terms containing the variational Beta-Liouville parameters once more.

$$\begin{aligned}
\mathcal{L}[\xi_q] = {} & \alpha_d(\Psi(\gamma_d)) - \Psi(\sum_{l=1}^{D} \gamma_l) + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma \\
& + \beta_\gamma)) + \beta(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N} \phi_n(\Psi(\gamma_l) - \Psi(\sum_{l=1}^{D} \gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N} \phi_{n(D+1)}(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& - (\log(\Gamma(\sum_{l=1}^{D} \gamma_l)) + \log(\gamma(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) \\
& - \log(\Gamma(\beta_\gamma)) - \log(\Gamma(\gamma_l)) \\
& + \gamma_l(\Psi(\gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) - \Psi(\sum_{l=1}^{D} \gamma_l) \\
& + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \beta_\gamma(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)))
\end{aligned} \tag{122}$$

To choose the words that involve Beta-Liouville variables with variations, including γ_i, α_γ, and β_γ, we obtain:

$$\begin{aligned}
\mathcal{L}(\gamma_i) = {} & \alpha_i(\Psi(\gamma_i)) - (\sum_{l=1}^{D} \alpha_l)(\Psi(\sum_{l=1}^{D} \gamma_l)) + \sum_{n=1}^{N} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D} \gamma_l)) \\
& - (\log \Gamma(\sum_{l=1}^{D}) - \log \Gamma(\gamma_i) + \gamma_i(\Psi(\sum_{l=1}^{D} \gamma_l) \sum_{d=1}^{D} \gamma_d)
\end{aligned} \tag{123}$$

and

$$\begin{aligned}\mathcal{L}[\alpha_\gamma] &= \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(-\Psi(\alpha_\gamma + \beta_\gamma)) \\ &+ (\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \sum_{n=1}^{N} \sum_{i=1}^{D} \phi_{ni} \sum_{n=1}^{N} \phi_{n(D+1)} (-\Psi(\alpha_\gamma + \beta_\gamma)) \\ &- (\log(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ &+ \beta_\gamma(-\Psi(\alpha_\gamma + \beta_\gamma)))\end{aligned} \tag{124}$$

Setting Eqs. 122, 123, and 124 to zero, we have the following update parameters:

$$\gamma_i = \alpha + \sum_{n=1}^{N} \phi_{ni} \tag{125}$$

$$\alpha_\gamma = \alpha + \sum_{n=1}^{N} \sum_{d=1}^{D} \phi_{nd} \tag{126}$$

$$\beta_\gamma = \beta + \sum_{n=1}^{N} \phi_{n(D+1)} \tag{127}$$

We address the issue of obtaining empirical Bayes estimates of the model parameters Υ and Ω by employing the variational lower bound as a substitute for the marginal log-likelihood, we maintain the variational parameters γ and Φ at the values obtained through variational inference. Afterwards, we calculate the empirical Bayes estimates by maximizing this lower bound in terms of the model parameters.

We obtain the equations for updating Ω_w. When we maximize Eq. 122 in relation to Ω, we obtain the subsequent equation:

$$\mathcal{L}[\Omega_w] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{D+1} \sum_{j=1}^{V} \phi_{dnl} w_{dn}^{j} \log(\Omega_{w(lj)}) + \sum_{l=1}^{D+1} \lambda_l (\sum_{j=1}^{V} \Omega_{w(lj)} - 1) \tag{128}$$

By calculating the derivative with respect to $\Omega_{w(lj)}$ and equating it to zero, we obtain:

$$\Omega_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \tau_{dnl} w_{dn}^{j} \tag{129}$$

The updates described below aim to reach a local maximum of a lower bound on $\log p(\Omega, \Upsilon|r)$, which is the best possible lower bound for any product approximations $q(\tau)q(w)$ of $p(\tau, w|\Omega, \Upsilon, r)$.

$$\Phi_l = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}\tau_{nl}(\lambda_n - 1)(\Psi(\gamma_l) - \Psi(\sum_{l=1}^{D}\gamma_l) \tag{130}$$

$$\gamma_l = \alpha_l + \sum_{n=1}^{N}\tau_{nl} \tag{131}$$

$$\Omega_{(lj)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}(\iota\sum_{d=1}^{M}\sum_{n=1}^{N_d}\tau_{dnl}w_{dn}^{j}) \tag{132}$$

Due to the fact that τ is defined in terms of the KL approximation, the variable Ω disappears in this scenario. In the second step, the algorithm now optimizes for τ. Since $q(w|\gamma, r, \tau)$ can be precisely modeled with multinomials, the minimum KL divergence is zero. As a result, the subsequent updates reach a local threshold of $\log p(\Omega, \tau|r)$

$$\gamma_l = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}\Omega\tau_{nl} \tag{133}$$

$$\tau_{nl} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}\Omega_{lv}e^{(\lambda_n-1)}e^{(\Psi(\gamma_i)-\Psi(\sum_{ii=1}^{D}\gamma_{ii})} \tag{134}$$

$$\Omega_{ij} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}(\iota + (\sum_{n}e^{(\lambda_n-1)}e^{(\Psi(\gamma_i)-\Psi(\sum_{ii=1}^{D}\gamma_{ii})}) \tag{135}$$

Considering that $\sum_{d=1}^{D+1}\phi_{n(d)} = 1$, for the normalization factor we have:

$$e^{\lambda_n-1} = \frac{1}{\tau_{(D+1)v}e^{(\lambda_n-1)}e^{(\Psi(\beta_\gamma)-\Psi(\alpha_\gamma+\beta_\gamma))} + \tau_{lv}e^{(\lambda_n-1)}e^{(\Psi(\gamma_i)-\Psi(\sum_{ii=1}^{D}\gamma_{ii}))}} \tag{136}$$

Variational Beta-Liouville for Word Level

In order to derive the update equations for the variational Bayesian learning, we follow the procedure of separating the terms that involve the variational Bayesian learning parameters.

$$\begin{aligned}
L[\xi_q] &= \alpha_d(\Psi(\gamma_d)) - \Psi(\sum_{l=1}^{D}\gamma_l) + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&+ \beta(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&+ \sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&+ \sum_{n=1}^{N}\phi_{n(D+1)}(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&- (\log(\Gamma(\sum_{l=1}^{D}\gamma_l)) + \log(\gamma(\alpha_\gamma+\beta_\gamma) - \log(\Gamma(\alpha_\gamma)) \\
&- \log(\Gamma(\beta_\gamma)) - \log(\Gamma(\gamma_i)) \\
&+ \gamma_i(\Psi(\gamma_i) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&- \Psi(\sum_{l=1}^{D}\gamma_l) + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)) \\
&+ \beta_\gamma(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma+\beta_\gamma)))
\end{aligned} \tag{137}$$

By choosing the terms that involve the variational Bayesian learning BL variables $\gamma_i, \alpha_\gamma, \beta_\gamma$, we have:

$$\begin{aligned}
L(\gamma_i) &= \alpha_i(\Psi(\gamma_i)) - (\sum_{l=1}^{D}\alpha_l)(\Psi(\sum_{l=1}^{D}\gamma_l)) + \sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l)) \\
&- (\log\Gamma(\sum_{l=1}^{D}) - \log\Gamma(\gamma_i) + \gamma_i(\Psi(\sum_{l=1}^{D}\gamma_l)\sum_{d=1}^{D}\gamma_d)
\end{aligned} \tag{138}$$

and

$$
\begin{aligned}
L[\alpha_\gamma] &= \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(-\Psi(\alpha_\gamma + \beta_\gamma)) \\
&+ (\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \sum_{n=1}^{N} \sum_{i=1}^{D} \phi_{ni} \\
&+ \sum_{n=1}^{N} \phi_{n(D+1)}(-\Psi(\alpha_\gamma + \beta_\gamma)) - (\log(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) + \alpha_\gamma(\Psi(\alpha_\gamma) \\
&- \Psi(\alpha_\gamma + \beta_\gamma)) + \beta_\gamma(-\Psi(\alpha_\gamma + \beta_\gamma)))
\end{aligned}
\tag{139}
$$

Taking the derivative of the above equations with respect to their Bayesian learning parameters gives:

$$
\begin{aligned}
\frac{\partial L[\gamma_i]}{\partial \gamma_i} &= \alpha_i \Psi^{'}(\gamma_i) - \Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{l=1}^{D} \alpha_l + \Psi^{'}(\gamma_i) \sum_{n=1}^{N} \phi_{ni} - D\Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{n=1}^{N} \phi_{ni} \\
&- (\Psi(\sum_{l=1}^{D} \gamma_l) + \gamma_i \Psi^{'}(\gamma_i) - \Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{d=1}^{D} \gamma_l - \psi(\sum_{l=1}^{D} \gamma_l))
\end{aligned}
\tag{140}
$$

and

$$
\begin{aligned}
\frac{\partial L[\gamma_i]}{\partial \alpha_\gamma} &= \alpha(\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) - \beta(\Psi^{'}(\alpha_\gamma + \beta_\gamma)) \\
&+ (\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) \sum_{n=1}^{N} \sum_{d=1}^{D} \phi_{nd} \\
&- \sum_{n=1}^{N} \phi_{n(D+1)}(\Psi^{'}(\alpha_\gamma + \beta_\gamma)) - (\alpha_\gamma(\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) \\
&- \beta_\gamma(\Psi^{'}(\alpha_\gamma + \beta_\gamma))
\end{aligned}
\tag{141}
$$

The variational BL update equations are obtained by setting the aforementioned equations to zero.

$$
\gamma_i = \alpha + \sum_{n=1}^{N} \phi_{ni} \tag{142}
$$

$$\alpha_\gamma = \alpha + \sum_{n=1}^{N}\sum_{d=1}^{D}\phi_{nd} \tag{143}$$

$$\beta_\gamma = \beta + \sum_{n=1}^{N}\phi_{n(D+1)} \tag{144}$$

Variational Parameter for Document Level

As we mentioned in Section 4.2.1, bi-RATM is a form of topic model that focuses on extracting each document's topic distribution for document embedding. ϵ_{2c+1} is interpreted as the attention signal for Φ_d, as per [37] description of the bi-RABP. $G_0 = \delta_{\Phi^d}$ is defined in the preceding variational inference and updates the topic distribution for the whole document, Φ^d, which represents the encoding of a single document. As detailed in the paper by [37], ϵ_{2c+1} is used as the attention signal for Υ_d. When $\{\Upsilon, \pi, \xi, \gamma\}$ During the alternating optimization process, these parameters are assumed constant, $\mathcal{L}(\tau, \Phi; \Upsilon, \Omega)$ is maximized with respect to Φ^d and Ω [37].

The following equation is derived by setting the derivative with respect to Ω to 0, where α_d, α, and β are initialized by BL [48].

$$\kappa = \alpha_d + \sum_{n=1}^{N}\phi_{ni} \tag{145}$$

$$\Upsilon = \alpha + \sum_{n=1}^{N}\sum_{d=1}^{D}\phi_{nd} \tag{146}$$

$$\varrho = \beta + \sum_{n=1}^{N}\phi_{n(D+1)} \tag{147}$$

According to Eqs. 145, 146, and 147, ν may be obtained as a normalized Ω:

$$\nu = \frac{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{148}$$

Beta-Liouville Parameters

The aim of this subsection is to compute the parameter estimates of the model using variational inference.

$$\begin{aligned} L[\xi] = & \sum_{m=1}^{M}(\log(\Gamma(\sum_{l=1}^{D}\alpha_l)) + \log(\Gamma(\alpha+\beta)) - \log\Gamma(\alpha) \\ & - \log\Gamma(\beta) - \sum_{i=1}^{D}\log\Gamma(\alpha_i) + \sum_{i=1}^{D}\alpha_i(\Psi(\gamma_{mi}) \\ & - \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) + \alpha(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}\beta_{m\gamma})) + \beta(\Psi(\beta_{m\gamma}) \\ & - \Psi(\alpha_{m\gamma}+\beta_{m\gamma}))) \end{aligned} \tag{149}$$

The derivative of the equation above with respect to the Bayesian learning parameter is expressed as:

$$\begin{aligned} \frac{\partial L[\xi]}{\partial \alpha_l} &= M(\Psi(\sum_{l=1}^{D}) - \Psi(\alpha_l)) + \sum_{m=1}^{M}(\Psi^{'}(\gamma_{ml}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) \\ \frac{\partial L[\xi]}{\partial \alpha} &= M[\Psi(\alpha+\beta) - \Psi(\alpha)] + \sum_{m=1}^{M}(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma})) \\ \frac{\partial L[\xi]}{\partial \beta} &= M[\Psi(\alpha+\beta) - \Psi(\beta)] + \sum_{m=1}^{M}(\Psi(\beta_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma})) \end{aligned} \tag{150}$$

The preceding equations clearly demonstrate that each BL parameter's derivative in Eq. 122 is influenced by not only its own value but also the values of the other Bayesian learning parameters. As a result, the optimization problem is addressed using the Newton-Raphson method. For this, we need to calculate the Hessian matrix in terms of the parameter space, as illustrated below:

$$\begin{aligned} \frac{\partial^2 L[\xi]}{\partial \alpha_l \alpha_j} &= M(-\delta(i,j)\Psi^{'}(\alpha_i) + \Psi^{'}(\sum_{l=1}^{D}\alpha_l)) \\ \frac{\partial^2 L[\xi]}{\partial \alpha^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\alpha)) \\ \frac{\partial^2 L[\xi]}{\partial \alpha \partial \beta} &= M\Psi^{'}(\alpha+\beta) \\ \frac{\partial^2 L[\xi]}{\partial \beta^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\beta)) \end{aligned} \tag{151}$$

By differentiating with respect to $\tau_{w(lj)}$ and equating it to zero, we obtain:

$$\tau_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \tag{152}$$

4.4 Experimental results

In this part, we demonstrate the effectiveness of our suggested techniques on two different and complex tasks, which are medical text topic modeling and text classification. We evaluate the performance of each model using standard metrics such as time complexity, log-likelihood, and perplexity [37, 10, 101]. Perplexity is a widely used measure in language modeling, and it is defined as follows:

$$prep(\mathcal{D}_{test}) = \exp\left(\frac{-\ln p(\mathcal{D}_{test})}{\sum_d |w_d|}\right) \tag{153}$$

where d is the length of the page and $|w_d|$ is its width. The average performance is better when the perplexity score is lower.

Comparison of the bi-RATM, GD-bi-RATM, BL-bi-RATM, LDA, and Seq-LDA performances is the major objective of both applications.

4.4.1 Topic Modeling for Medical Text

The objective of text classification is to assign documents to one or more predetermined subject categories [102]. Much research has been done on this problem, and many solutions have been proposed [103, 46]. In natural language processing, topic modelling is one of the most used methods. Topic models can be used for a wide range of applications, from analyzing different kinds of texts like news articles and tweets to making graphs of related topics and documents.

Topic modeling is an interesting technique for dealing with problems that have high dimensionality and sparsity, like health and medical text mining. However, despite the abundance of data available, there is still a requirement to improve the effectiveness of this method [107]. This method was first shown to analyze text, with documents as the objects and the number of times a phrase

was used as the feature. The term ”topic modelling” generally refers to a set of statistical learning techniques used to uncover latent topics in a large corpus of text data, without the need for supervision.

Therefore, a topic refers to a combination of keywords that follows a probability distribution, and a document consists of a combination of topics, also following a probability distribution. It is worth noting that a topic model only provides a set of keywords for each topic, according to [13]. Although topic modeling is a useful technique for mining health and medical text, there is still room for improvement given the vast amount of available data [107].

To evaluate our models, we chose the medical transcription dataset [130], mental health dataset [131], Genia dataset [132], nematode biology abstract and TMVAr corpus from the PubMed website [111].

4.4.2 Medical Transcription Dataset

Medical data is challenging to obtain due to the Health Insurance Portability and Accountability Act (HIPAA) privacy regulations. However, the MTSamples dataset presents a remedy by providing samples of medical transcriptions.

MTSamples dataset provides access to a large library of transcribed medical reports for a wide range of medical specialties and employment types. These example reports are offered exclusively for reference purposes by various transcriptionists and users.

Table 4.1 shows the result for the GD-bi-RATM for the topics, and Fig. 4.1 shows the time complexity for the bi-RATM, GD-Bi-RATM and BL-Bi-RATM models. Also, Table 4.2 compares the perplexity of mentioned approaches, and all the results illustrate that the BL-bi-RATM algorithm outperforms the medical transcription dataset.

Mental Health Dataset

Our mental health is influenced by our emotional health as well as our psychological and social well-being. Therefore, a healthy mental state is necessary in order to live a balanced and healthy existence. It affects the way that we think, feels, and behave as a result. In addition, it influences how we respond to stressful events, interact with others, and make ultimate decisions. Emotional

Table 4.1: Common topics identified with BL-bi-RATM model in the Medical Transcript dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'normal', 'skin', 'incision', 'stable', 'preoperative', 'using', 'performed', 'point', 'pulmonary', 'evidence'
Topic 2	'history', 'patient', 'blood', 'removed', 'artery', 'right', 'weight', 'pressure', 'obtained', 'tissue'
Topic 3	'patient', 'wound', 'history', 'abdominal', 'abdomen', 'general','closed', 'surgery', 'bleeding', 'surgery'
Topic 4	'laparoscopic", 'placed', 'incision', 'removed', 'performed', 'approximately', 'normal', 'yearold', 'brought', 'femoral'
Topic 5	'right', 'history', 'performed', 'patient', 'lower', 'anterior', 'procedure', 'pulmonary', 'heart', 'present'

Table 4.2: Comparison for the perplexity for different models, indicating model fit quality across different topic numbers (K) on the MT dataset

K	5	10	15	20
Seq-LDA	-2437.02	-2218.03	-2197.86	-1769.49
bi-RATM	-335229.04	-326538.03	-314397.13	-302863.32
GD-bi-RATM	-1341231.01	-1274679.18	-1362211.57	-1195597.23
BL-bi-RATM	-5422819.15	-4957706.67	-4119113.31	-3846521.56

and mental health are significant since they affect ideas, habits, and emotions and are vital to life [131]. Being emotionally healthy can increase productivity and effectiveness in tasks such as jobs, school, and caregiving. Maintaining good mental health is crucial for healthy relationships, as it helps to cope with life changes and difficulties. Although mental health issues are common, there is help available, and people with mental illnesses can recover. A dataset containing frequently asked questions (FAQs) related to mental health is used to validate our models [131].

The BL-bi-RATM algorithm's top 5 topics are shown in Table 4.3, and a comparison of other methods based on perplexity is presented in Table 4.4. According to the tables, the BL-based model has the lowest perplexity among all the tested models, indicating superior performance in this regard.

Fig. 4.2 displays the time complexity of three algorithms bi-RATM, GD-bi-RATM and BL-bi-RATM. We can conclude from the figure that the BL-based bi-RATM has the lowest time complexity.

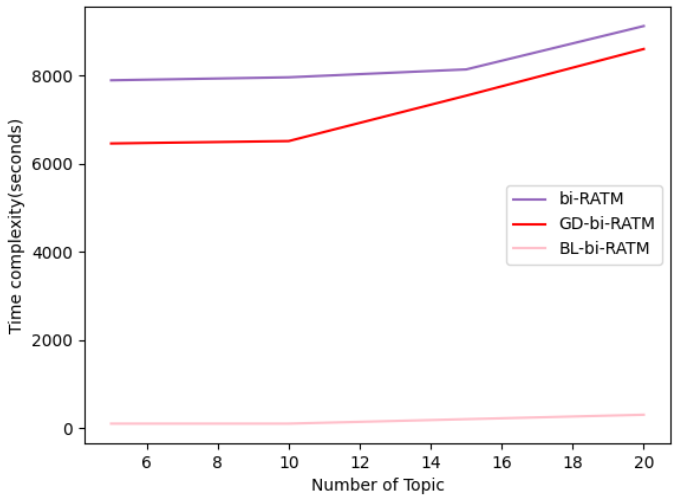


Figure 4.1: Time Complexity for MT dataset (min)

Genia Dataset

Biomedical texts contain a wealth of information that can be applied to medical advancements. Previously, domain specialists would manually extract such information. Automating this information extraction operation can aid in field progress. As an illustration, a biomedical text might demonstrate the impact of medications on a person. They can also be used to diagnose medical issues in people. As a result, automated event extraction from biomedical texts is quite advantageous. It includes the original biomedical text, labeled trigger words, the location of the trigger word inside the text, and the event type associated with the trigger word. Table 4.5 shows the result for the BL-bi-RATM for the topics, and Table 4.6 illustrate the BL-bi-RATM model has less perplexity compared to other models. Also, Fig. 4.3 shows the time complexity of bi-RATM, GD-bi-RATM and BL-bi-RATM models.

4.4.3 Topic Modeling

Topic models are frequently used in document clustering and organizing text data collections. These models can also assist in the classification of text [104].

Given the volume of documents, it is inefficient to analyze each one manually. Instead, one technique is identifying the terms that best characterize the corpus, such as the most frequent words.

Table 4.3: Common topics identified with GD-bi-RATM model in the Mental health dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'family', 'health', 'disorder', 'support', 'mental', 'anxiety', 'depression', 'loved', 'child', 'selfhelp'
Topic 2	'health', 'depression', 'people', 'information', 'mental', 'think', 'physical', 'illness', 'problems', 'thinking'
Topic 3	'symptoms', 'people', 'disorder', 'different', 'loved', 'information','problem', 'health', 'friends', 'illness'
Topic 4	'mental', 'people', 'health', 'services', 'disorder', 'things', 'important', 'young', 'learn', 'support'
Topic 5	'important', 'mental', 'support', 'health', 'learn', 'people', 'feelings', 'illness', 'anxiety', 'different'

Table 4.4: Comparison of the perplexity for different models, indicating model fit quality across different topic numbers (K) on the Mental health dataset

K	5	10	15	20
Seq-LDA	-1968.56	-1886.64	-1768.47	-1589.54
bi-RATM	-4531.27	-4389.31	-4299.87	-4120.03
GD-bi-RATM	-5896.65	-5370.23	-4961.31	-4512.12
BL-bi-RATM	-17811.21	-13708.67	-12517.36	-12718.17

Another approach would be to break the documents into words and phrases and then organize the words and phrases into groups according to similarity. Then, the generated word and phrase clusters can be used to gain a more profound knowledge of the corpus. Intuitively, the corpus is the collection of words chosen by selecting one from each category. The rule-based text mining techniques that utilize regular expressions and the dictionary-based keyword searching strategies differ from topic modelling. Instead, it attempts to identify the essential words or subjects in a text corpus without prior knowledge [13].

In order to verify the effectiveness of our proposed models, we selected a set of 2246 documents from the Associated Press [13].

The results for topic selection and perplexity of each model for this dataset are shown in Tables 4.9 and 4.10. According to the results, BL-bi-RATM and GM-based bi-RATM models have a smaller perplexity. Furthermore, as Fig. 4.5 illustrates, the time complexity values using the proposed BL-based model are smaller than other models.

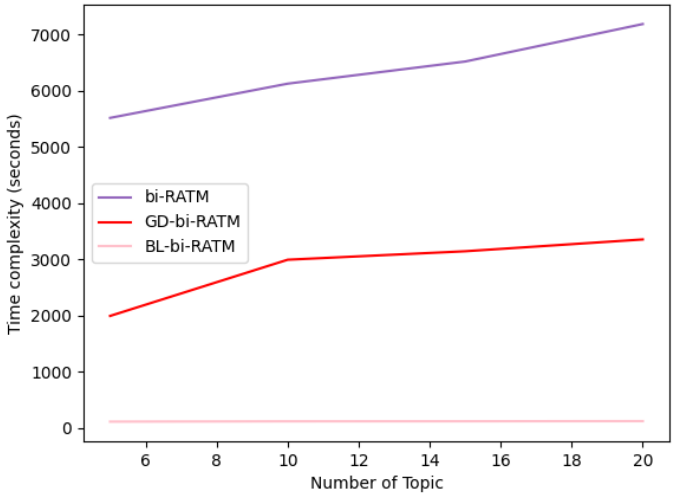


Figure 4.2: Time complexity for dataset

We also applied the mentioned models (Seq-LDA, bi-RATM, GD-bi-RATM, BL-bi-RATM) in another dataset, namely the "CMU Book Summary Dataset" [105], to validate our proposed model. The dataset consists of plot summaries for 16,559 books collected from Wikipedia, along with aligned metadata from Freebase, which includes information such as author, title, and genre.

The top 5 topics for the GD-bi-RATM approaches are presented in Table 4.11, and the success rates of using these models on the dataset are shown in Table 4.12. As per the results, it is observed that the GD-bi-RATM outperforms the other models in this case. Fig. 4.6 displays the time complexity of the three models under different training conditions.

Table 4.5: Common topics identified with BL-bi-RATM model in the Genia dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	expression', 'activation', 'transcription', 'cells', 'nuclear', 'binding', 'activity', 'analysis', 'kappa', 'results'
Topic 2	'human', 'cells', 'activity', 'nfkappab', 'activation', 'kappa', 'protein', 'nfkappa', 'expression', 'factor'
Topic 3	activation', 'expression', 'factor', 'positiveregulation', 'transcription', 'human', 'promoter', 'nfkappa', 'binding', 'geneexpression'
Topic 4	'binding', 'protein', 'activation', 'mediated', 'region', 'induced', 'monocytes', 'level', 'sites', 'function'
Topic 5	'cells', 'transcription', 'nfkappa', 'activation', 'proteins','induced', 'transcriptional', 'activity', 'nuclear', 'specific'

Table 4.6: Comparison of the perplexity for different models, indicating model fit quality across different topic numbers (K) on the Genia dataset

K	5	10	15	20
Seq-LDA	-989.89	-902.03	-823.92	-751.97
bi-RATM	-1001.98	-984.11	-843.65	-752.23
GD-bi-RATM	-1107.98	-1089.13	- 972.63	-892.31
BL-bi-RATM	-1261.91	-1150.29	-1021.15	-958.99

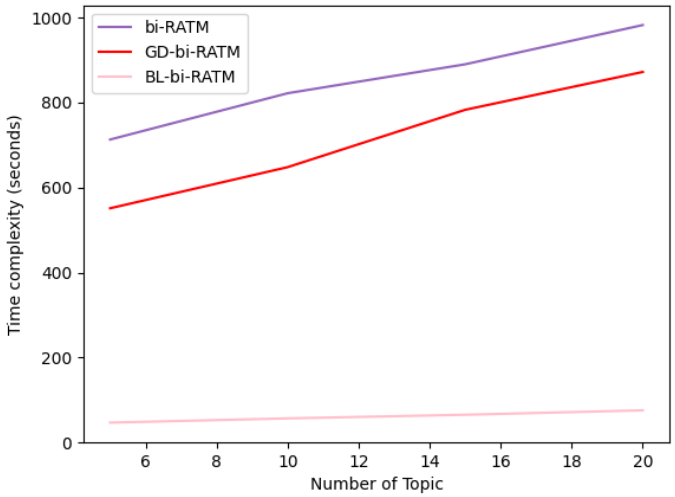


Figure 4.3: Time complexity for Genia dataset

Table 4.7: Common topics identified with BL-bi-RATM model in tmVar dataset, each defined by a set of keywords t

Topic No	Topics
Topic 1	‘novel’, ‘associated’, ‘mutation’, ‘dnamutation’, ‘patients’, ‘deletion’, ‘genes’, ‘allele’, ‘chinese’, ‘results’
Topic 2	‘proteinmutation’, ‘variants’, ‘allele’, ‘analysis’, ‘polymorphisms’, ’expression’, ‘results’, ‘family’, ‘patient’, ‘mutation’
Topic 3	‘patients’, ‘mutations’, ‘genes’, ‘results’, ‘mutant’, ‘deletion’, ’mutation’, ‘proteinmutation’, ‘identified’, ‘cells’
Topic 4	‘polymorphism’, ‘dnamutation’, ‘mutations’, ‘genetic’, ‘association’, ’proteinmutation’, ‘analysis’, ‘deletion’, ‘genotype’, ‘patient’
Topic 5	‘patients’, ‘mutations’, ‘study’, ‘compared’, ‘nucleotide’, ‘allele’, ‘proteinmutation’, ‘dnamutation’, ‘genetic’, ‘codon’

Table 4.8: Comparison of the perplexity for different models, indicating model fit quality across different topic numbers (K) on tmVar dataset

K	5	10	15	20
Seq-LDA	-210.67	-201.51	-184.34	-175.67
bi-RATM	-253.98	-236.33	-204.98	-192.02
GD-bi-RATM	- 321.17	-287.71	-259.31	-218.92
BL-bi-RATM	-441.21	-314.78	-292.28	-249.15

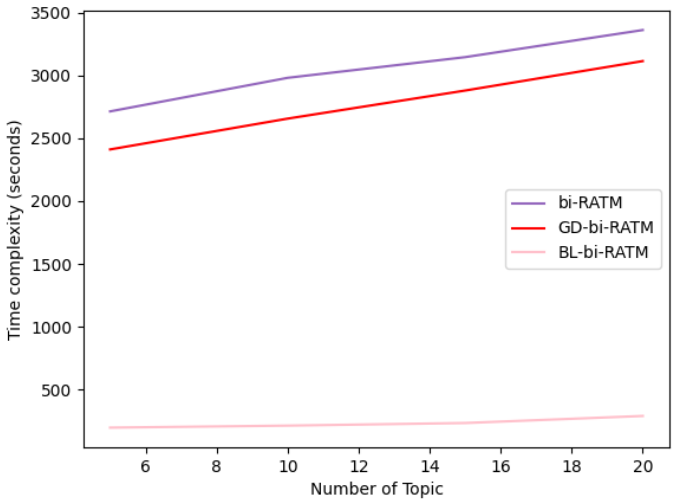


Figure 4.4: Time complexity for tmVar dataset

Table 4.9: Common topics identified with GD-bi-RATM model in the Associated press dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'million', 'percent', 'billion', 'state', 'government', 'company', 'international', 'workers', 'department', 'president'
Topic 2	'percent', 'government', 'including', 'program', 'police', 'united', south', 'political', 'service', 'party'
Topic 3	'house', 'states', 'police', 'money', 'business', 'federal', 'soviet', including', 'lower', 'allowed'
Topic 4	'federal', 'market', 'american', 'percent', 'later', 'million', president', 'think', 'billion', 'increase'
Topic 5	'police', 'workers', 'percent', 'officials', 'state', 'minister', official', 'group', 'called', 'government'

Table 4.10: Comparison of the perplexity for different models, indicating model fit quality across different topic numbers (K) on Associated press dataset

K	5	10	15	20
Seq-LDA	-1249.67	-1195.67	-1120.07	-1006.91
bi-RATM	-12114.54	-11953.36	-10034.66	-1489.68
GD-bi-RATM	-2954197.21	-2703086.19	-2311242.12	-1986589.45
BL-bi-RATM	-167289.44	-135154.30	-121314.45	-111032.89

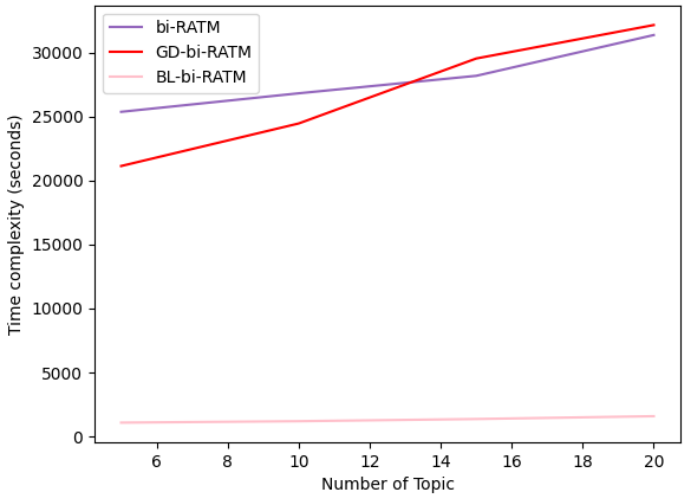


Figure 4.5: Time complexity for Associated Press dataset

Table 4.11: Common topics identified with BL-bi-RATM model in the CMU Book dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'people', 'states', 'united', 'percent', 'family', 'soviet', 'government', 'president', 'germany', 'country'
Topic 2	'percent', 'soviet', 'american', 'yearold', 'police', 'government', 'including', 'million', 'expected', 'people'
Topic 3	'defense', 'federal', 'south', 'military', 'department', 'house', 'called', 'children', 'union', 'national'
Topic 4	'million', 'people', 'chief', 'market', 'black', 'troops', 'committee', 'earlier', 'government', 'department'
Topic 5	'state', 'people', 'million', 'united', 'billion', 'years', 'campaign', 'national', 'prices', 'nations'

Table 4.12: Comparison of the perplexity for different models, indicating model fit quality across different topic numbers (K) on the CMU Book dataset

K	5	10	15	20
Seq-LDA	-20001.92	-19685.05	-19031.11	-18329.57
bi-RATM	-21567.33	-20891.59	-20091.02	-18979.59
GD-bi-RATM	-29847.19	-25551.35	-22386.53	-21663.15
BL-bi-RATM	-259174.91	-225062.91	-221131.02	-218795.39

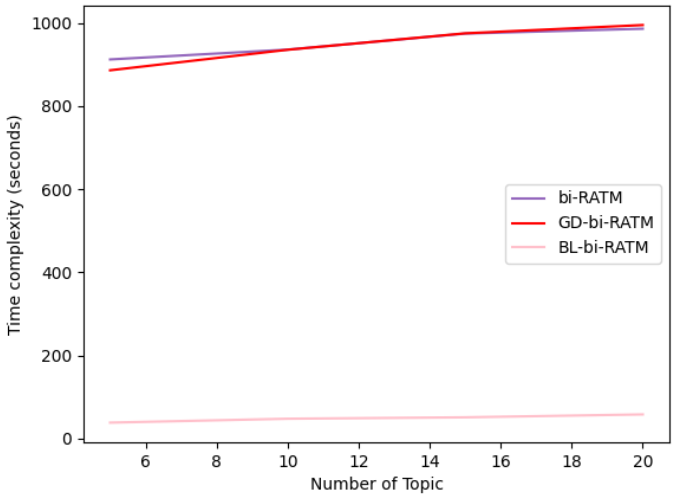


Figure 4.6: Time complexity for CMU dataset

Chapter 5

Flexible Distribution Approaches to Enhance Regression and Deep Topic Modelling Techniques

5.1 Introduction

The past two decades have seen an explosion of digital content, resulting in an unprecedented amount of text data generated daily. This surge has created a need for efficient and accurate text analysis technologies. Machine learning and deep learning have become powerful tools for different type of applications [5, 9, 133]

Text mining and topic modeling are vital for processing the vast amounts of text data generated daily. The Bag of Words (BoW) model [120] simplifies text structure for easier algorithmic processing. Topic modeling, a subfield of text mining, identifies themes in documents, with latent Dirichlet allocation (LDA) [13] using BoW to learn topics by assuming documents are mixtures of topics characterized by word distributions. Topic models are split into generative models, like LDA, and matrix decomposition techniques, like probabilistic latent semantic analysis (pLSA) [11, 12], which uses maximum likelihood to find topics. LDA improves on pLSA by treating topic mixtures as Dirichlet variables, offering a robust generative model [13].

Many collections of documents come with additional information, such as metadata [134] and annotations. For instance, a book might have information about the author, or an article could have tags describing its topic. Images may be included with product reviews, and clinical records might have structured information about the patient. These annotations can help guide the learning of topic models [135]. Incorporating this extra information into the topic model is possible using either downstream [136] or upstream models [137].

By incorporating additional information, downstream and upstream models can be used to improve topic modeling accuracy [136, 138]. Downstream models, including supervised LDA [66], use annotations to guide the topic model learning process directly. For example, in the analysis of product reviews, downstream models may incorporate information such as ratings, comments, and images associated with the reviews [71, 67]. Meanwhile, upstream models, such as Dirichlet multinomial regression (DMR) [67], use annotations to preprocess the data before topic modeling. An example of upstream modeling is using named entity recognition to identify entities within the documents, then using these entities as additional features in the topic modeling process. Both downstream and upstream models can enhance the accuracy of topic modeling and provide deeper insights into the underlying themes and patterns in the document collection [139].

Although DMR is a flexible approach to incorporating document features, it is often limited to a few features. There are several reasons for this. First, many text corpora have a limited number of document-level features available. Second, as the dimensionality of the model grows, the hyperparameters become increasingly difficult to interpret. Finally, when the dimensionality of the document features is high, DMR is prone to overfitting the hyperparameters [71]. As a result, in practice, DMR is typically applied in settings with a limited number of features or where the analyst hand-selects a few relevant features. Despite these limitations, DMR remains a powerful tool for analyzing text data, offering an efficient approach to topic modeling that can be adapted to various applications[67].

A possible solution to DMR's document feature limitation is to use low-dimensional representations of those features. In recent years, neural networks have shown exceptional learning success in generalizable representations, which can eliminate the need for manually designed features [140]. Furthermore, neural networks can handle different data types, including text, images, and

other metadata features, making them suitable for addressing dimensionality reduction in DMR [57, 141]. To that end, the deep Dirichlet multinomial regression (dDMR) model has been proposed [71]. This model extends DMR by incorporating a deep neural network that can learn a transformation of the input metadata into features that are then used to form the Dirichlet hyperparameter. By jointly learning a feature representation for each document and a log-linear function that captures the distribution over topics, dDMR provides a powerful approach for modeling complex relationships between document features and topics [71].

DMR and dDMR topic modeling are powerful tools in natural language processing for analyzing text data; however, both have limitations [71, 136]. The standard Dirichlet distribution [79] assumes that the response variables are independent, which is often not the case in real-world applications. Additionally, the Dirichlet distribution cannot model over- or underdispersion in the data. These limitations can result in poor model fit and inaccurate predictions. To address these issues, we propose an extension of DMR and dDMR using the collapsed Gibbs sampling algorithm [142] with two alternative distributions, namely, the GD distribution [64] and the Beta-Liouville distribution [65].

In this chapter, we present our proposed models, namely generalized Dirichlet multinomial regression (GDMR), deep generalized Dirichlet multinomial regression (dGDMR), Beta-Liouville multinomial regression (BLMR), and deep Beta-Liouville multinomial regression (dBLMR). Firstly, we describe the characteristics of the fitting distribution associated with each proposed model. Next, we use Gibbs sampling to estimate the parameters of each distribution. Finally, we provide the complete learning algorithm for our models.

5.2 The Considered Distributions

5.2.1 Generalized Dirichlet Multinomial Regression

The generalized Dirichlet distribution was introduced in [90], and its covariance structure is more extensive than that of the Dirichlet distribution. The GD distribution solves the Dirichlet distribution's restrictions, which include the assumptions of negative correlation and equal confidence. Thus, it has become a suitable option as a prior in Bayesian learning settings [91, 143].

The GDMR assigns the probability mass for a count vector $X = (x_1, \ldots, x_d)$ given a parameter set $\xi = (\alpha_1, \ldots, \alpha_{d-1}, \beta_1, \ldots, \beta_{d-1})$ where all α_i, β_i values are positive, and it is described in [91, 144].

$$\begin{aligned} \mathcal{GDM}(\mathbf{X}|\xi) &= \binom{m}{X} \prod_{i=1}^{d-1} \frac{\Gamma(\alpha_i + x_i)}{\Gamma(\alpha_i)} \frac{\Gamma(\beta_i + z_{i+1})}{\Gamma(\beta_i)} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i + \beta_i + z_i)} \\ &= \binom{m}{X} \prod_{i=1}^{d-1} \frac{(\alpha_i)_{x_i} (\beta_i)_{z_{i+1}}}{(\alpha_i + \beta_i)_{z_i}} \end{aligned} \tag{154}$$

where $z_i = \sum_{l=i}^{d} x_l$ is the cumulative sum.

To connect the covariates $\mathbf{X}$ to the parameters, [144] employed the subsequent link functions: $\alpha_i = e^{y^T \alpha_i}$, and $\beta_i = e^{y^T \beta_i}$. Assuming that the parameter set $\xi = \{\alpha, \beta\}$ includes all the regression coefficients, the log-likelihood is described in [144].

$$\begin{aligned} \mathcal{L}_n(\mathcal{X}|\xi) = \sum_{j=1}^{n} ln \binom{m_j}{X_j} + \sum_{i=1}^{d} \sum_{j=1}^{n} \Bigg(\sum_{k=0}^{x_{ij}-1} ln(e^{y_j^T \alpha_i} + k) + \\ \sum_{k=0}^{z_{i,j+1}-1} ln(e^{y_j^T \beta_i}) - \sum_{k=0}^{z_i - 1} ln(e^{y_j^T \alpha_i} + e^{y_j^T \beta_i} + k) \Bigg) \end{aligned} \tag{155}$$

where k is altered from j to z_{ij}, and $z_{ij} = \sum_{i=1}^{n} \sum_{l=i}^{d} x_{il}$.

The GDMR is better suited for modeling count data than the DMR, as it has a more flexible covariance structure and additional parameters. Several studies have been conducted on both distributions, as mentioned in the literature [91, 144, 145].

Link Functions for Generalized Multinomial Dirichlet Distribution

For the GD distribution, the parameter $\vartheta = \{\alpha_i, \beta_i\}$ can be linked to the p-dimensional covariates vector X as:

$$\alpha_i = \lambda_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p) \tag{156}$$

$$\beta_i = \lambda_2(\beta_i x_1 + \beta_i x_2 + \ldots + \beta_i x_p), \qquad i = 1, \ldots, d \tag{157}$$

For finding $\rho(\mu_i)$, the following procedure has been followed:

$$\rho(\mu_i) = X_i^T \vartheta, \qquad i = 1, \ldots, d \tag{158}$$

then we have:

$$\rho_1(\mu_i) = X_i^T \alpha_i \tag{159}$$

$$\rho_2(\mu_i) = X_i^T \beta_i \tag{160}$$

Hence, for the remainder of this paper, for the GDMR, we symbolize a_i as $e^{x_j^T \alpha_i}$ and b_i as $e^{x_j^T \beta_i}$.

5.3 Generalized Dirichlet Multinomial Regression Topic Modeling

Previously, we mentioned that the GD distribution, much like the Dirichlet distribution, serves as a conjugate prior to the multinomial distribution. However, the GD has a more comprehensive covariance matrix compared to the Dirichlet distribution, as stated in [91]. Due to this characteristic, we will be utilizing the collapsed Gibbs sample approach to establish an extension to DMR founded on the assumption of the GD.

To generate the document representation, a vector X is employed in d dimensions, encapsulating vital metadata values as characteristics. If a metadata attribute denotes the existence or non-existence of chosen features, the respective elements in x_d will be allocated a value of 1 for every specified feature, whereas other locations will hold a value of 0. Further, the model integrates a default feature set at 1 to accommodate for the average value of each topic [67].

The generative process is detailed below, where X embodies the data matrix, $\mathcal{N}$ indicates the normal distribution, $\mathcal{GD}$ symbolizes the GD distribution, and $\mathcal{M}$ represents the multinomial distribution. A graphical representation of the GDMR topic model is shown in Fig. 5.1.

(1) For any topic t:

  (a) Draw $m \sim \mathcal{GD}(\xi)$.

(2) For any d documents:

(a) $\rho = \rho(\mu_i)$

(b) Draw $\Omega \sim \mathcal{GD}(\rho)$.

(c) for each word i:

i. Draw $z_i \sim \mathcal{M}(m)$

ii. Draw $w_i \sim \mathcal{M}(\Omega)$

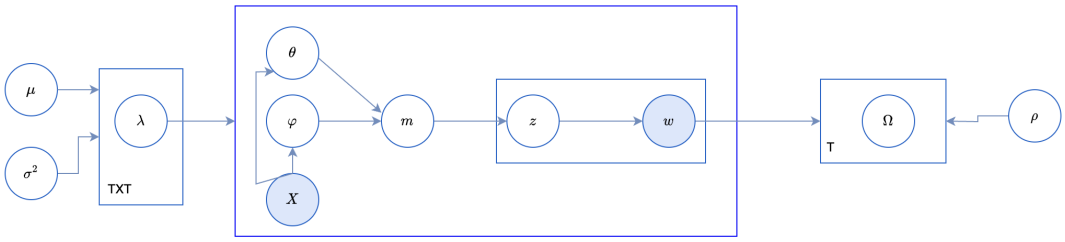


Figure 5.1: Graphical representation of "upstream" GDMR model

$GD(\xi)$ is a GD distribution in d dimensions, where the parameters are denoted by ξ and are represented as $(a_1, b_1, \ldots, a_d, b_d)$. The probability distribution function is indicated by p. The variable Λ_i is calculated as the difference between b_i, a_{i+1}, and b_{i+1} ($\Lambda_i = b_i - a_{i+1} - b_{i+1}$) [46]:

$$p(m_1, \ldots, m_d|\xi) = \prod_{i=1}^{d} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} m_i^{a_l - 1}(1 - \sum_{j=1}^{i} m_j)^{\Lambda_i} \tag{161}$$

.

With our GD prior available, we move forward by establishing the word topic probability matrix, denoted as Ω. By assuming the variable's conditional independence, we can derive the joint distribution as follows:

$$p(m, z, w, |\xi, \Omega) = p(m|\xi)p(w|z, \Omega)p(z|m) \tag{162}$$

where z denotes the set of latent topics.

Integrating over the m parameters and the topic space gives the following:

$$p(w|\xi,\Omega) = \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)} \int m^{a_i-1}(1-\sum_{j=1}^{i} m_j)^{\Lambda_i} \times \prod_{n=1}^{N}\prod_{j=1}^{d+1}\prod_{j=1}^{V}(m\Omega_{ij})^{w_n^j} d\theta \tag{163}$$

In Eq. 163, ξ and Ω are the corpus-level parameters and m is a document-level parameter, while z and w are word-level parameters.

5.3.1 Inference via Collapsed Gibbs Sampling

Based on the generative process of the GD distribution in 5.6.1 the full generative equation $p(X, z, \theta, \varphi, w|, \Omega, \xi, \mu)$ of our new approach is also expressed as:

$$p(X, z, \theta, \varphi, w|, \Omega, \xi, \mu) = p(w|\mu)p(\theta|\Omega)p(\varphi|\xi) \times \prod_{n=1}^{N} p(z_n|\theta)p(x_n|z_{nn}, \varphi) \tag{164}$$

The GD document prior distribution with hyperparameter $\Omega = (\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n)$ is denoted by $p(\theta|\Omega)$, while the GD corpus prior distribution with hyperparameters $\xi = (a_1, b_1, \ldots, a_d, b_d)$ is denoted by $p(\varphi|\xi)$. The Bayesian inference process involves approximating the posterior distribution of the latent variables z, after marginalizing out the parameters:

$$p(X, z|w, \Omega, \xi) = W \int_{\theta}\int_{\varphi} p(X, z, \theta, \varphi, |\Omega, \xi) d\varphi d\theta \tag{165}$$

It is significant to note that the joint distribution can be represented as a product of Gamma functions, as indicated in previous studies [13, 95, 64]. This allows for the formulation of the expectation for the true posterior distribution, as shown below:

$$p(z_{ij} = k|X, w, \Omega, \xi) = \mathrm{E}_{p(z^{-ij}|w,X,\Omega,\xi)}[p(z_{ij} = k|z^{-ij}, X, w, \Omega, \xi)] \tag{166}$$

Thus, using the GD prior, the posterior is computed as follows:

$$p(z_{ij}=k|z^{-ij},X,w,\Omega,\xi) \propto \left[\frac{(N_{jk}^{-ij}+a_{wk})(b_{wk}+\sum_{l=k+1}^{K+1}N_{jl}^{-ij})}{(a_{wk}b_{wk}+\sum_{l=k+1}^{K+1}N_{jl}^{-ij})}\right] \times \left[\frac{(N_{kv_{ij}}^{-ij}+\alpha_v)(\beta_v+\sum_{d=v}^{V+1}N_{kd_{ij}}^{-ij})}{\alpha_v+\beta_v+\sum_{d=v}^{V+1}N_{kd_{ij}}^{-ij})}\right] = A(K) \tag{167}$$

where N^{ij} refers to counts where the superscript $-ij$ indicates the associated variables excluding x_{ij} and z_{ij}.

Normalizing the aforementioned distribution now yields a posterior probability denoted as:

$$p(z_{ij}=k|z^{-ij},x,\Omega,\xi) = \frac{A(k)}{\sum_{k'=1}^{K}A(k')} \tag{168}$$

From Eq. 166 to Eq. 168, the collapsed Gibbs sampling computes the entire sampling procedure as follows:

$$p(z_{ij}=k|X,w,\Omega,\xi) = \mathrm{E}_{p(z^{-ij}|w,X,\Omega,\xi)}\left[\frac{A(k)}{\sum_{k'=1}^{K}A(k')}\right] \tag{169}$$

The use of collapsed Gibbs sampling in our GD-based model allows us to sample from the actual posterior distribution p, as shown in Eq. 169. This sampling method is more precise compared to the approach used in variational-based models, where samples are typically drawn from an estimated distribution [96, 64]. Therefore, we can conclude that our model is more accurate.

After sampling, the parameter estimation process uses the techniques presented in [96, 97]. Next, the empirical likelihood method [96] is used to validate the results for the held-out dataset. This process ultimately leads to the estimation of the class conditional probability $p(X|w,\Omega,\xi)$ within the collapsed Gibbs sampling framework:

$$p(X|w,\Omega,\xi) = \prod_{ij}\sum_{k=1}^{K}\frac{1}{S}\sum_{s=1}^{S}\tilde{\theta}_{jks}\tilde{\varphi}_{kws} \tag{170}$$

The parameters are subsequently calculated as:

$$\tilde{\theta}_{jks} = \frac{(N_{jk} + a_{wk})(b_{wk} + \sum_{l=k+1}^{K+1} N_{jl})}{(a_{wk} b_{wk} + \sum_{l=k+1}^{K+1} N_{jl})} \tag{171}$$

$$\tilde{\varphi}_{kws} = \frac{(N_{kv_{ij}} + \alpha_v)(\beta_v + \sum_{d=v}^{V+1} N_{kd_{ij}})}{\alpha_v + \beta_v + \sum_{d=v}^{V+1} N_{kd_{ij}})} \tag{172}$$

where S is the size of sample.

5.4 Deep Generalized Dirichlet Multinomial Regression

Our method enhances the GDMR by using a deep neural network to transfer input metadata into features that comprise the GD hyperparameter. Unlike DMR models that use a log-linear function of document features for document-topic priors, our proposed model, which we call deep generalized Dirichlet multinomial regression (dGDMR), simultaneously learns a feature representation for each document as well as a log-linear function that accurately captures topic distribution. Furthermore, because the neural network is in charge of mapping document features to topic priors, we use gradient ascent and back-propagation [146, 147] to improve performance by optimizing both the topic model and the neural network parameters.

DGDMR replaces the log-linear model used in GDMR with an arbitrary function denoted f that maps a real-valued vector with dimension F to a K-dimensional representation. We make no assumptions about the precise form of this function, focusing instead on minimizing the output cost through gradient ascent optimization. In actual implementation, we employ a neural network tailored to the particular category of document characteristics, such as a convolutional neural network for images. We prefer neural networks because they can express intricate functions, generalize well to new data, and facilitate joint training via gradient ascent and back-propagation.

The vocabulary size is denoted as V and the number of topics as K. It is worth noting that in practical applications, the document features do not need to be limited to fixed-length feature vectors. For example, the function f could be implemented as a recurrent neural network that maps from a sequence of characters to a fixed-length vector in $\mathbb{R}^K$. The GDMR model is actually a special

case of our proposed model, dGDMR, where the function f is chosen to be linear.

In dGDMR, the log-linear model is supplanted by an arbitrary function referred to as f. This function transforms a real-valued vector of dimension F into a representation of dimension K. The selection of neural networks is favored due to their potent expressive capabilities, their capacity to adapt effectively to unobserved data, and their unique advantage in joint training. Below we show the generative process of dGDMR [71].

(1) Define the $f \in \mathbb{R}^F \rightarrow \mathbb{R}^K$.

(2) Define the topic-word prior parameters $w \in \mathbb{R}^V$.

(3) Generate document priors for each document m with features $\theta_m \in \mathbb{R}^F$:

  (a) Draw $\tilde{\theta}_m = exp(f(\alpha_m, \beta_m))$.

  (b) $\theta_m \sim GD(\tilde{\theta}_m)$.

(4) Word distribution generated for each topic K:

  (a) $\tilde{\phi}_k = exp(w)$

  (b) $\tilde{\phi}_k \sim GD(\tilde{\phi}_k)$

(5) Generate data for each token (m, n)

  (a) Unobserved topic: $z_{mn} \sim \theta_m$

  (b) Observed word: $w_{mn} \sim \phi_{z_{mn}}$

5.4.1 Parameter Estimation

The random variables of the topic model are inferred using collapsed Gibbs sampling, and the model parameters are estimated using gradient ascent with back-propagation. To maximize the log-likelihood of token and topic assignments, we use alternating optimization: one iteration of collapsed Gibbs sampling (sample topics for each word) and then an update of the parameters of f by gradient ascent. According to Eq. 155, log-likelihood can be expressed as follows, where $\tilde{\theta}_{\alpha,k} = e^{y_j^T \alpha_i}$ and $\tilde{\theta}_{\beta,k} = e^{y_j^T \beta_i}$:

$$\mathcal{L}_n(\mathcal{X}|\xi) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\left(\sum_{k=0}^{x_{ij}-1} ln(\tilde{\theta}_{\alpha,k}+k) + \sum_{k=0}^{z_{i,j+1}-1} ln(\tilde{\theta}_{\beta,k}) - \sum_{k=0}^{z_i-1} ln(\tilde{\theta}_{\alpha,k}+\tilde{\theta}_{\beta,k}+k)\right) \tag{173}$$

The sampling step in dGDMR remains the same as in GDMR when the parameters are given. We estimate the network parameters by employing back-propagation through the network for a fixed sample. The gradient of the data log-likelihood, denoted as L, is expressed in Eq. 173, where ψ represents the digamma function (the derivative of the log-gamma function), n_m denotes the number of tokens in document m, and $n_{m,k}$ represents the count of tokens that were assigned to topic k in document m.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\tilde{\theta}} &= \psi(\sum_{k=1}^{K}\tilde{\theta}_{\alpha,k}) + \psi(\sum_{k=1}^{K}\tilde{\theta}_{\beta,k}) - \psi(\sum_{k=1}^{K}\tilde{\theta}_{\beta,k}\sum_{k=1}^{K}\tilde{\theta}_{\alpha,k}+n_m) \\ &+ \psi(\tilde{\theta}_{\beta,k}+\tilde{\theta}_{\alpha,k}+n_m) - \psi(\tilde{\theta}_{\beta,k}) - \psi(\tilde{\theta}_{\beta,k})\end{aligned} \tag{174}$$

5.5 Multinomial Beta-Liouville Regression

As outlined in [148], the Liouville family of distributions, specifically of the second kind, encompasses the Dirichlet distribution as a specific instance, given certain conditions. Notably, the Dirichlet distribution is seen as a particular case of the Beta-Liouville if $\beta_d = \alpha_{d+1} + \beta_{d+1}$ [**?**, 46]. These conditions involve maintaining the same normalized variance for all components in the Liouville random vector and employing a Beta distribution as the variate generating the density. The Beta-Liouville distribution is the term given when the Beta distribution is used for generating density, as articulated in [149]. Like the Dirichlet distribution, the Beta-Liouville distribution can serve as a conjugate prior to the multinomial distribution. However, it also has the ability to overcome the Dirichlet distribution's main limitations. Furthermore, the Beta-Liouville distribution has two additional parameters that can be used to adjust its spread, making it more practical and allowing

for better modeling capabilities. Using the Beta-Liouville distribution as a prior for the multinomial distribution yields a flexible joint distribution known as the multinomial Beta-Liouville (MBL) distribution, [95, 93].

5.6 Beta-Liouville Multinomial Regression Topic Modeling

The Beta-Liouville multinomial regression (BLMR) model is established with the same underlying scenario, where a collection of documents is assumed to follow a similar pattern as described in the GDMR section. This allows us to capture the relationships between documents and topics, while taking into account the additional parameters and flexibility offered by the Beta-Liouville distribution. By leveraging the properties of the Beta-Liouville distribution within the framework of multinomial regression and topic modeling, the BLMR model provides a unique and versatile approach for analyzing and understanding document collections in a wide range of applications.

A multinomial prior w is chosen for a particular topic, representing the probability that each term in the vocabulary is associated with that topic. This prior, denoted by $\Omega_{w_{ij}}$, is used to calculate the probability that a given word, w^j, is assigned to a specific topic, z^i, i.e., $p(w^j = 1|z^i = 1)$. First, the syllables are sampled at random following this multinomial prior. Then, using a multinomial probability distribution, the probability of observing a specific word, w_n, given its topic assignment, z_n, and the multinomial prior Ω_w is calculated as $p(w_n|z_n, \Omega_w)$.

In addition, the topic assignment distribution, denoted by $BL(\Upsilon)$, follows a d-variable Beta-Liouville distribution with parameters $\Upsilon = (\eta_1, ..., \eta_D, \eta, \tau)$. The probability distribution function of this distribution can be expressed as:

$$P(\theta_1, \ldots, \theta_D|\Upsilon) = \frac{\Gamma(\sum_{d=1}^{D}\eta_d)\Gamma(\eta+\tau)}{\Gamma(\eta)\Gamma(\tau)}\prod_{d=1}^{D}\frac{\theta_d^{\eta_d-1}}{\Gamma(\eta_d)}\times \left(\sum_{d=1}^{D}\theta_d\right)^{\eta-\sum_{l=1}^{D}\eta_l}\times\left(1-\sum_{l=1}^{D}\theta_l\right)^{\tau-1} \tag{175}$$

The mean, variance, and covariance are used to depict the distribution of the Beta-Liouville

distribution.

$$E(\theta_d) = \frac{\eta}{\eta + \tau} \frac{\eta_d}{\sum_{d=1}^{D} \eta_d} \tag{176}$$

$$\begin{aligned} var(\theta_d) = (\frac{\eta}{\eta + \tau})^2 \frac{\eta_d(\eta_d + 1)}{(\sum_{m=1}^{D} \eta_m)(\sum_{m=1}^{D} \eta_m + 1)} \\ - E(\theta_d)^2 \frac{\eta_d^2}{(\sum_{m=1}^{D} \eta_m)^2} \end{aligned} \tag{177}$$

and the covariance between θ_l and θ_k is given by:

$$Cov(\theta_l, \theta_k) = \frac{\eta_l \eta_k}{\sum_{d=1}^{D} \eta_d} \left(\frac{\frac{(\eta+1)(\eta)}{(\eta+\tau+1)(\eta+\tau)}}{\sum_{d=1}^{D} \eta_d + 1} - \frac{\frac{\eta}{\eta+\tau}}{\sum_{d=1}^{D} \eta_d} \right) \tag{178}$$

In the preceding equation, the covariance matrix of the Beta-Liouville distribution is more inclusive than the covariance matrix of the Dirichlet distribution.

5.6.1 Proposed Link Functions for MBL Regression

We can express the relationship between the parameters and the p-dimensional covariate vector $Y = (y_1, \ldots, y_p)$ in the following forms for regression based on the MBL distribution [93]:

$$\begin{aligned} \eta_i &= g_1(\eta_i y_1 + \eta_i y_2 + \ldots + \eta_i y_p), \qquad i = 1, \ldots, d \\ \eta &= g_2(\eta y_1 + \eta y_2 + \ldots + \eta y_p), \\ \tau &= g_3(\tau y_1 + \tau y_2 + \ldots + \tau y_p) \end{aligned} \tag{179}$$

To determine the value of $g(\mu_j)$, the following approach is employed:

$$g(\mu_j) = Y_j^T \Upsilon \qquad j = 1, \ldots, n \tag{180}$$

where μ_j represents the average of Y_j, while Υ denotes a vector of regression parameters. As such, the relationship can be expressed as follows:

$$logit(\mu_j) = \log\Big(\frac{\mu_j}{1-\mu_j}\Big), \tag{181}$$

and for $logit$ link function we have the following:

$$\Pi_j(y) = \frac{\exp(\Upsilon^T Y_j)}{1+\sum_{j=1}^{n-1}\exp(\Upsilon^T Y_j)} \tag{182}$$

Therefore, in the case of the MBL model, we have the following:

$$\begin{aligned} g_1(\mu_j) &= Y_j^T \eta_i \\ g_2(\mu_j) &= Y_j^T \eta \\ g_3(\mu_j) &= Y_j^T \tau \end{aligned} \tag{183}$$

Therefore, from the rest of this paper, for the BLMR, we symbolize η_i as $e^{x_j^T \alpha_i}$, η as $e^{x_j^T \alpha}$ and τ as $e^{x_j^T \beta}$.

Considering the parameter set Υ as including all the regression coefficients, denoted as $(\alpha_1, \ldots, \alpha_{d-1}, \alpha, \beta)$, the complete log-likelihood can be expressed as follows:

$$\begin{aligned} \mathcal{L}_n(\mathcal{X}|\theta) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\Bigg[& \sum_{k=0}^{x_{ij}-1} ln(\alpha_i + k) \\ & + \sum_{k=0}^{z_{ij}} ln(\beta + k) + \sum_{k=0}^{x_{i-1}} ln(\alpha + k) \\ & - \sum_{k=0}^{x_{i,m}-1} ln(\alpha_i + k) - \sum_{k=0}^{x_{i+1}} ln(\alpha + \beta + k)\Bigg] \end{aligned} \tag{184}$$

The generative procedure is described as follows, where X stands for the data matrix, $\mathcal{N}$ designates the normal distribution, $\mathcal{BL}$ signifies the Beta-Liouville distribution, and $\mathcal{M}$ typifies the multinomial distribution. An illustrative diagram of the BLMR topic model can be seen in Fig. 5.2:

(1) For any topic t:

(a) Draw $m \sim \mathcal{BL}(\Upsilon)$.

(2) For any d documents:

(a) $g = g(\mu_i)$

(b) Draw $\Omega \sim \mathcal{BL}(g)$.

(c) for each word i:

i. Draw $z_i \sim \mathcal{M}(m)$

ii. Draw $w_i \sim \mathcal{M}(\Omega)$

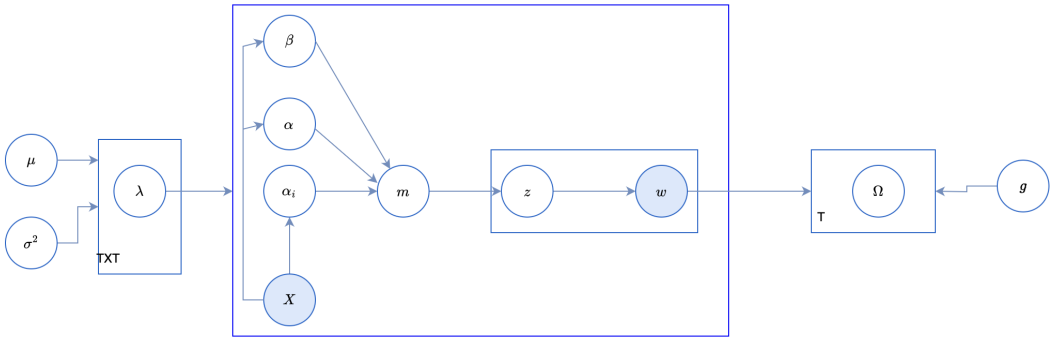


Figure 5.2: Graphical representation of "upstream" BLMR model

5.6.2 Inference via Collapsed Gibbs Sampling

The Collapsed Gibbs Sampler (CGS) contributes to inference by estimating posterior distributions through a Bayesian network of conditional probabilities, which are determined via a sampling process of hidden variables. When compared to the traditional Gibbs sampler that functions in the combined space of latent variables and model parameters, the CGS offers significantly faster estimation. As CGS operates in the collapsed space of latent variables, in the joint distribution $p(X, z, \theta, \phi, w|\Omega, \Upsilon, \mu)$, the model parameters θ, ϕ are marginalized out to obtain the marginal joint distribution $p(X, z, w|\Omega, \Upsilon, \mu)$ defined as:

$$p(x, z, w|\Omega, \Upsilon) = W \int_{\theta} \int_{\varphi} p(X, z, \theta, \varphi, w|\Omega, \xi) d\varphi d\theta \tag{185}$$

By employing Eq. 185, the method estimates the conditional probabilities of the latent variables z_{ij} based on the current state of all variables, while excluding the individual variable z_{ij} [97]. The collapsed Gibbs sampler, on the other hand, estimates the topic assignments for the observed words by utilizing the conditional probability of latent variables, where $-ij$ refers to counts or variables with z_{ij} removed [97]. This particular conditional probability is defined as [100]:

$$p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon) = \frac{p(z_{ij}, z^{-ij}, X, w|\Omega, \Upsilon)}{p(z^{-ij}, X, w|\Omega, \Upsilon)} \tag{186}$$

The sampling mechanism of the collapsed Gibbs approach can be summarized as an expectation problem:

$$p(z_{ij} = k|X, w, \Omega, \Upsilon) = \mathrm{E}_{p(z^{-ij}|w,X,\Omega,\Upsilon)}[p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon)] \tag{187}$$

The collapsed Gibbs sampling Beta-Liouville multinomial procedure consists of two phases for assigning documents to clusters. First, each document is assigned a random cluster for initialization. After that, each document is assigned a cluster based on the Beta-Liouville distribution after a specified number of iterations.

The goal is to use a network of conditional probabilities for individual classes to sample the latent variables from the joint distribution $p(X, z|w, \Omega, \Upsilon)$. The assumption of conjugacy allows the integral in Eq. 185 to be estimated.

$$p(X, z|w, \upsilon) = C \prod_{j=1}^{M} \left[\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)\Gamma(\alpha + \beta)}{\prod_{i=1}^{k} \Gamma(\alpha_i)\Gamma(\alpha)\Gamma(\beta)} \right] \times \frac{\prod_{i=1}^{k} \Gamma(\alpha_i')\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha' + \beta')\Gamma(\sum_{i=1}^{K} \alpha_i')} \tag{188}$$

The likelihood of the multinomial distribution, defined by the parameter Υ, and the probability

density function of the Beta-Liouville distribution can be expressed as follows:

$$
\begin{aligned}
p(X|\Upsilon) &= \int p(X|\theta)p(\theta|\alpha_1,\ldots,\beta,\alpha)d\theta \\
&= \int \prod_{k=1}^{k} \theta_k^{m_k} \frac{\Gamma(\sum_{k=1}^{k}\alpha_k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{k=1}^{K} \frac{\theta_k^{\alpha_k-1}}{\Gamma(\alpha_k)} \\
&\times (\sum_{k=1}^{K}\theta_k)^{\alpha-\sum\alpha_k}(1-\sum_{k=1}^{K}\theta_k)^{\beta-1}d\theta
\end{aligned}
\tag{189}
$$

By integrating the probability density function of the Beta-Liouville distribution over the parameter θ and incorporating updated parameters derived from the remaining integral in Eq. 191, we are able to express it as a fraction of Gamma functions. The following shows the updated parameters, where N_{jk} represents counts corresponding variables. [95, 100]:

$$
\begin{aligned}
\alpha_K^{'} &= \alpha_k + \sum_{j=1}^{k} N_{jk} \\
\alpha^{'} &= \alpha + N_{jk} \\
\beta^{'} &= \beta + N_{jk}
\end{aligned}
\tag{190}
$$

The Eq. 189 is then equivalent to:

$$
\begin{aligned}
p(k|\alpha_1,\ldots,\alpha_k,\beta,\alpha) = &\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{k=1}^{k-1}m_k)\Gamma(\beta+m_k)}{\Gamma(\alpha)\Gamma(\beta)\prod_{k=1}^{K}\Gamma(\alpha_k)\Gamma(\sum_{k=1}^{K}(\alpha_k+m_k))} \\
&\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+m_k)}{\Gamma(\alpha+\sum_{k=1}^{K-1}m_k+\beta+m_k)}
\end{aligned}
\tag{191}
$$

The parameters $\alpha_1,\ldots,\alpha_k,\alpha$, and β correspond to the Beta-Liouville distribution, while m_k represents the number of documents in cluster k.

After the sampling process, parameter estimation is performed. Subsequently, the empirical likelihood method [96] is utilized to validate the results using a held-out dataset. Ultimately, this process leads to the estimation of the class conditional probability $p(X|w,\Omega,\Upsilon)$ within the framework of collapsed Gibbs sampling:

$$p(X|w,\Omega,\Upsilon) = \prod_{ij}\sum_{k=1}^{K}\frac{1}{S}\sum_{s=1}^{S}\tilde{\theta}_{jks}\tilde{\varphi}_{kws} \tag{192}$$

The parameters are then computed as follows:

$$\tilde{\theta}_{jks} = \frac{(N_{jk}+\alpha_k)(\alpha_{jk}+\sum_{l=k+1}^{K+1}N_{jl})(N_{jk}+\beta_k)}{(a_k b_k+\sum_{l=k+1}^{K+1}N_{jl})(\alpha_j+\sum_{l=k+1}^{K+1}N_{jl})} \tag{193}$$

$$\tilde{\varphi}_{kws} = \frac{(N_{jk}+\alpha_w)(\alpha_{jw}+\sum_{l=k+1}^{K+1}N_{jl})(N_{jk}+\beta_w)}{(\alpha_w b_w+\sum_{l=k+1}^{K+1}N_{jl})(\alpha_{wj}+\sum_{l=k+1}^{K+1}N_{jl})} \tag{194}$$

where S is the size of sample.

5.7 Deep Beta-Liouville Multinomial Regression

Our proposed method, deep Beta-Liouville multinomial regression (dBLMR), improves on BLMR by converting input metadata into features for the Beta-Liouville hyperparameter using a deep neural network. Like GDMR, it simultaneously learns a feature representation for each document and a log-linear function for accurate topic distribution. The topic model and neural network parameters are optimized using gradient ascent and back-propagation. The deep version, dGDMR, replaces the log-linear model with an arbitrary function denoted as f and optimizes it using gradient ascent. Because of their ability to express complex functions, generalize well, and facilitate joint training, neural networks are preferred.

5.7.1 Inference

The assumptions for BLMR are similar to those for GDMR, with vocabulary size denoted as V and the number of topics denoted as K. It is also worth noting that document features can be more adaptable in practical applications than fixed-length feature vectors. In our proposed topic model, the random variables are inferred using collapsed Gibbs sampling, which is a Markov chain-Monte Carlo (MCMC) method commonly used for topic modeling. This sampling method permits us to estimate each document word's posterior distribution of topic assignments. Gradient ascent

with back-propagation, a prominent optimization algorithm for training neural networks, is used to estimate the model parameters.

An alternating optimization strategy is implemented to maximize the log-likelihood of token and topic assignments. Each iteration includes one round of collapsed Gibbs sampling to generate topic samples for each document word. This enables us to revise the topic assignments based on the current model parameter estimations. Gradient ascent is then used to update the parameters of the function f by computing the gradients of the log-likelihood with respect to the parameters and updating them accordingly. This procedure of alternating optimization is repeated until convergence is achieved.

Using a combination of collapsed Gibbs sampling, gradient ascent, and back-propagation, our model can effectively learn both the topic assignments and the parameters of the function f in a mutually advantageous technique. The Gibbs sampling phase assists in refining the topic assignments based on the current estimate of the parameters. At the same time, the gradient ascent step modifies the parameters of f to capture the underlying patterns in the data more accurately. This joint optimization strategy improves the performance of our model by maximizing the log-likelihood of token and topic assignments. According to Eq. 184 , the following equation holds if $\tilde{\zeta}_{\alpha_i} = e^{y_j^T \alpha_i}$, $\tilde{\zeta}_{\beta} = e^{y_j^T \beta}$ and $\tilde{\zeta}_{\alpha} = e^{y_j^T \alpha}$:

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\zeta) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\Bigg[\sum_{k=0}^{x_{ij}-1}(\tilde{\zeta}_{\alpha_i}+k) \\ + \sum_{k=0}^{z_{ij}}(\tilde{\zeta}_{\beta}+k) + \sum_{k=0}^{x_{i-1}}(\tilde{\zeta}_{\alpha}+k) \\ - \sum_{k=0}^{x_{i,m}-1}(\tilde{\zeta}_{\alpha_i}+k) - \sum_{k=0}^{x_{i+1}}(\tilde{\zeta}_{\alpha}+\tilde{\zeta}_{\beta}+k)\Bigg]\end{aligned} \tag{195}$$

In our proposed model, dBLMR, the sampling phase remains the same when parameters are fixed. Therefore, we estimate the network parameters by back-propagating a fixed sample through the network. The data log-likelihood gradient, denoted by L, is mathematically expressed in Eq. 195, where ψ represents the digamma function, the derivative of the log-gamma function.

$$\frac{\partial \mathcal{L}}{\partial \tilde{\zeta}} = \psi(\sum_{k=1}^{K} \tilde{\zeta}_{\alpha_i,k}) + \psi(\sum_{k=1}^{K} \tilde{\zeta}_{\beta,k}) + \psi(\sum_{k=1}^{K} \tilde{\zeta}_{\alpha,k}) - \psi(\sum_{k=1}^{K} \tilde{\zeta}_{\alpha_i,k} + \sum_{k=1}^{K} \tilde{\zeta}_{\beta,k} + \sum_{k=1}^{K} \tilde{\zeta}_{\alpha,k} + n_m) + \psi(\tilde{\zeta}_{\alpha_i,k} + \tilde{\zeta}_{\beta,k} + \tilde{\zeta}_{\alpha,k} + n_m) - \psi(\tilde{\zeta}_{\alpha_i,k}) - \psi(\tilde{\zeta}_{\beta,k}) - \psi(\tilde{\zeta}_{\beta,k}) \tag{196}$$

5.8 Experimental Results

In this section, we demonstrate the efficacy of our suggested methodologies by experimenting with two complex tasks: medical text topic modeling and text classification. Standard evaluation metrics such as time complexity, log-likelihood, and perplexity are commonly used in language modeling to evaluate the performance of our models [37, 10, 101]. Perplexity is generally defined as follows, where the page length, denoted as d, and its width, represented by $|w_d|$, are factors in our model. Lower perplexity scores indicate adequate average performance.

$$prep(\mathcal{D}_{test}) = \exp\left(\frac{-\ln p(\mathcal{D}_{test})}{\sum_d |w_d|}\right) \tag{197}$$

5.8.1 Topic Modeling for Medical Texts

The primary goal of text classification is to systematically assign various documents into one or more previously determined subject categories, as elucidated by previous research [102]. This area has been extensively studied, and a multitude of potential solutions have been suggested [103, 46]. Topic modeling, a prevalent method in the field of natural language processing, has emerged as an exceptionally effective strategy for this purpose. It is versatile and finds applications in the analysis of various types of texts such as news articles and tweets, and even in the creation of graphical representations of interconnected topics and documents.

Topic modeling serves as a fascinating approach for addressing problems characterized by high dimensionality and sparsity, which are common in areas like health and medical text mining. Despite the vast amount of data currently accessible, there remains a necessity to enhance the efficacy

of this method [107]. Initially, this approach was developed to analyze text, where documents served as the subject matter and the frequency of phrase usage constituted the feature.

The phrase ”topic modeling” refers to a collection of statistical learning techniques aimed at discovering hidden or ’latent’ topics within a sizable body of text data, all without requiring explicit supervision. Thus, in the context of topic modeling, a ’topic’ signifies a set of keywords that adhere to a probability distribution, and a ’document’ comprises a blend of such topics, also abiding by a similar probability distribution.

According to [13], it is important to note that a topic model merely provides a set of keywords corresponding to each topic. Topic modeling proves instrumental for mining health and medical text, contributing to better understanding and extraction of useful insights from such data. However, given the enormous volume of data available, there is a clear need for continued improvements and advancements in topic modeling techniques [107].

In the domain of BioNLP [150], or biological natural language processing, topic modeling carries notable advantages. It allows for the processing and understanding of complex and domain-specific biological texts, facilitating more efficient information extraction and enabling more targeted and efficient research. Furthermore, in the context of biological data, topic modeling could reveal hidden thematic structures or latent topics, which could potentially unearth new correlations and insights in the field of biology. This also aids in the categorization and organization of large volumes of biological data, making it more accessible and easier to analyze for researchers and professionals in the field.

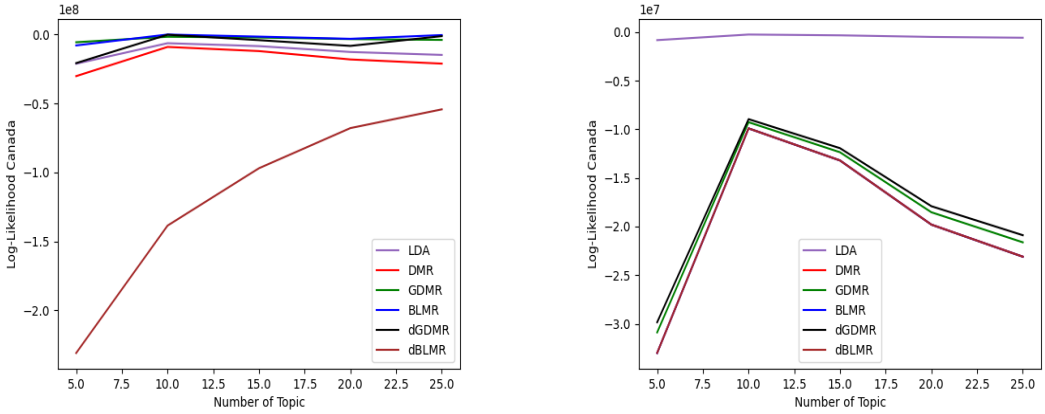
To evaluate our models, we chose the Covid-Tweet, Mental Health Tweet, Symptom for Disease, and Drugs Side Effects metadata.

Covid-Tweet

This dataset originates from Twitter, a renowned social media platform. It comprises a variety of tweets which have been manually labeled for the purpose of facilitating text classification. Privacy is of paramount importance and hence, identifiable information such as names and usernames have been anonymized using unique codes. The structure of the data includes four key features: location, which provides the geographical location of the Twitter user; Tweet Atthat, which shows when

the tweet was posted; Original Tweet, which has the content of the tweet itself; and Label, the assigned tag or category for the tweet. To validate the proposed models, we decided to focus on the 'Location' feature as the basis for our analysis. We aimed to uncover hidden topic models based on the geographical location of the tweets. This approach can help in detecting location-specific trends, sentiments, or patterns which could provide insight for various applications. Specifically, we have demonstrated this by focusing on tweets from New York City and Canada, and performing topic modeling on these subsets of data. The resulting tables detail our findings, illustrating the unique topics and patterns we discovered within the tweets originating from these two geographical locations.

The prediction results for this dataset are shown in Tables 5.1, 5.2, Fig. 5.3 and Fig. 5.4. We can see from their relatively higher perplexity that LDA and DMR are not the best fitted to the data. Furthermore, DBLMR has a lower perplexity; we can thus conclude that a dBLMR-based regression model is better for Covid-Tweet dataset.

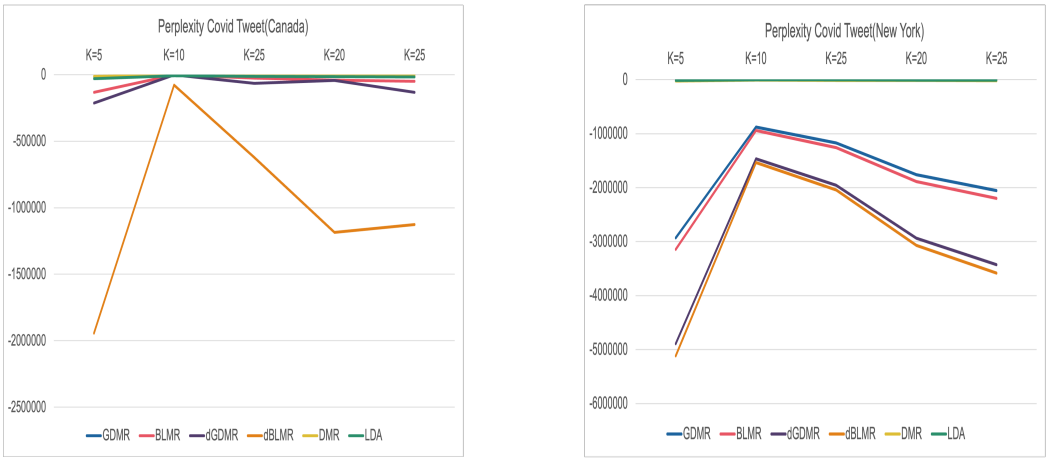


(a) Canada (b) New York

Figure 5.3: Log-likelihood comparison for Covid Tweet

Mental Health Tweet

The Mental Health Corpus represents a collection of textual data, pertaining specifically to individuals suffering from various mental health conditions such as anxiety and depression. The



(a) Canada (b) New York

Figure 5.4: Perplexity comparison for Covid Tweet

structure of this corpus is twofold: one column is dedicated to the comments or discussions surrounding mental health issues, and the other serves as a categorical indicator, labeling whether these comments are deemed toxic or not. This dataset provides an ample ground for a broad range of analytical applications including sentiment analysis, detection of harmful or toxic language, and the study of language patterns associated with mental health discussions. The information in this corpus can serve as a valuable resource for diverse stakeholders such as researchers, mental health professionals, and any individual keen on delving deeper into the discourse and sentiments related to mental health matters. In our research, we focused our experimental analysis on determining the presence or absence of mental health conditions based on the text data from the Mental Health Corpus. Our methodology involved implementing topic modeling to extract hidden patterns or themes from the dataset. Topic modeling is a type of statistical modeling used for discovering the abstract ”topics” that occur in a collection of documents. In this context, these ”documents” are the comments or discussions in the Mental Health Corpus. Through this technique, we aimed to extract meaningful and significant information from the textual data, thereby enhancing our understanding of the language dynamics surrounding mental health issues.

The results for topic selection and perplexity of each model for this dataset are shown in Table 5.3 and Fig. 5.6. The results show that dGDMR and dBLMR models have a smaller perplexity. Furthermore, as Table 5.4 illustrates, the time complexity values using the proposed BL-based model

Table 5.1: Common topics identified with BLMR model in the Canada and New York subcategories , each defined by a set of keywords

Topic	Canada	New York
Topic1	'starting"grocery', 'turbo', 'positive', 'pandemic', 'working', 'people 'covid', 'coronavirus' , 'consumer'	coronavirus', 'grocery', 'negative' 'prices', 'store', 'crisis', 'retail', 'spending', 'consumer', 'additive',
Topic2	'prices', 'negative', 'small', 'store', 'positive', 'workers', 'retail', 'working', 'customers', 'crisis'	'sanitizer', 'grocery', 'consumer', 'prices', 'online', 'stores', 'spending' 'coronavirus', 'behavior', 'economy',
Topic3	'covid', 'grocery', 'customers,' 'situation','prices', 'farmers', 'turbo', 'coronavirus', 'positive' , 'consumer'	'negative', 'store', 'covid', 'grocery', 'lysol', 'retail','pandemic' 'coronavirus', 'business', 'amazon'
Topic4	'coronavirus', 'positive', 'grocery', 'customers', 'march', 'customers', 'march', 'providing','negative', 'online', 'store'	'consumer', 'positive', 'store', 'sanitizer', 'retail', 'online', 'americans', 'retail', 'online', 'americans', 'covid','coronavirus', 'disinfectant'
Topic5	'positive', 'covid', 'store', 'business', 'demand', 'working', 'online', 'turbo', 'consumer', 'march'	'supermarket', 'store', 'consumer', , 'covid', 'coronavirus', , 'positive' 'online', 'laundry', 'additive', 'negative'

Table 5.2: Time complexity comparison for different model at varying topic levels (K) on subdatsset Canada (min)

K	5	10	15	20	25
LDA	45	58.5	63	67.5	71.1
DMR	23	29.9	32.2	34.5	36.34
GDMR	11.83	24.843	28.392	30.758	31.941
BLMR	9.53	22.543	26.092	28.458	29.641
dGDMR	255.04	280.544	357.056	408.064	433.568
dBLMR	251.64	327.132	352.296	377.46	397.5912

are smaller than other models.

Symptom for Disease

This dataset comprises 1200 entries, each attributed to a specific disease label and accompanied by a descriptive text detailing the associated symptoms in natural language. The 'label' column signifies the disease while the 'text' column provides symptom descriptions pertinent to that disease. The data incorporates 24 distinct diseases, each represented by 50 unique symptom descriptions. Therefore, we have an equal distribution of data points for all diseases, which totals up to 1200 data points. The spectrum of diseases covered in this dataset is broad, encompassing conditions such as psoriasis, varicose veins, typhoid, chicken pox, impetigo, dengue, fungal infection, common cold,

Table 5.3: Common topics identified with d BLMR model in the Mental Health Tweet dataset, each defined by a set of keywords

Topic No	Topics
Topic1	'azarkansero', 'youre', 'autism', 'difference', 'sleep', 'depression' 'michaelsos', 'watch', 'learn', 'better'
Topic2	'depression', 'genevieveverso', 'misslusyd', 'really', 'right 'think', 'overcome', 'amazing', 'sosdaily'
Topic3	"overcome', 'depression', 'happened', 'friend', 'treatments', 'making', 'start', 'break', 'story', 'birthday'
Topic4	'thefuxedos', 'happy', 'overcome', 'birthday', 'misslusyd', 'friends', 'treatments', 'thinking', 'school', 'autism'
Topic5	'people', 'depression', 'misslusyd', 'great', 'world', 'comes', 'therapy', 'treatments', 'really', 'support'

Table 5.4: Time complexity comparison for different model at varying topic levels (K) on Mental Health Tweet dataset (min)

K	5	10	15	20	25
LDA	142.32	185.01	199.24	213.48	224.86
DMR	135.67	176.37	189.93	203.50	214.35
GDMR	120.26	156.33	168.36	180.39	190.01
BLMR	119.03	154.73	166.64	178.54	188.06
dGDMR	137.06	150.76	191.88	219.29	233.00
dBLMR	135.86	176.68	190.20	203.79	214.65

pneumonia, dimorphic hemorrhoids, arthritis, acne, bronchial asthma, hypertension, migraine, cervical spondylosis, jaundice, malaria, urinary tract infection, allergy, gastroesophageal reflux disease, drug reaction, peptic ulcer disease, and diabetes. Our analytical approach with this dataset revolves around the 'label' feature, that represents each disease. We employed topic modeling to uncover hidden themes or patterns related to each disease based on the symptom descriptions provided in the 'text' column. Topic modeling is an unsupervised machine learning method that discovers abstract themes or "topics" from a collection of documents. Here, these "documents" correspond to the symptom descriptions for each disease. This strategy helps us in deciphering and elucidating the language and patterns associated with each disease, thus enhancing our understanding of symptom descriptions and their implications in the field of medical text analysis.

The top 5 topics for the dBLMR approaches are presented in Table 5.5, and the perplexity of using these models on the dataset are shown in Fig. 5.8. As per the results, it is observed that the

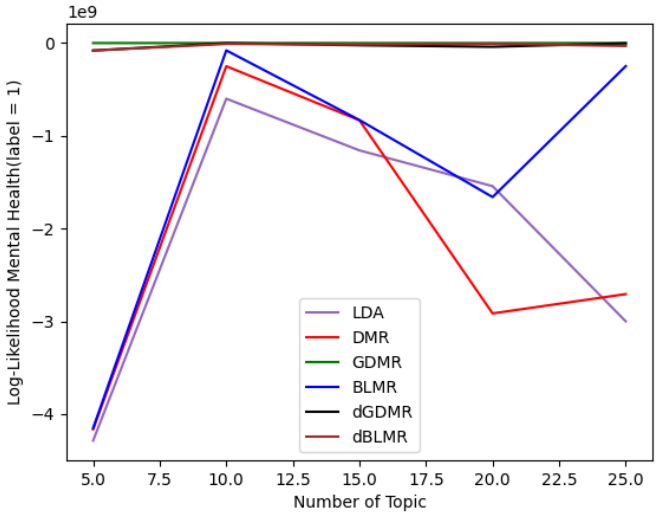


Figure 5.5: Log-likelihood for Mental Health Tweet dataset

dBLMR outperforms the other models in this case. The figure presented in Fig. 5.7 displays the log-likelihood of the five models under different training conditions.

Table 5.5: Common topics identified with GDMR model in the Symptom for Disease (Cancer) dataset, each defined by a set of keywords

Topic No	Topics
Topic1	radiation, 'effects', 'vomiting', 'breathing', 'temporary', 'malignant', 'cramps', 'problems', 'hands', 'blood'
Topic2	'unnamed', 'sores', 'chills', 'malignant', 'cancer', 'mouth', 'uncontrolled', 'fluorouracil', 'growth', 'names'
Topic3	thiotepa, medical, 'trouble', 'disease', 'throat', 'resulting', 'growth', 'cancer', 'vomiting', 'balance '
Topic4	'effects', 'tumour', 'permanently', 'growth', 'carcinoma', 'nausea', 'unusual', 'short', 'names', 'trouble'
Topic5	'bruising', 'unusual', 'bleeding', 'breath', 'growth', 'urine', 'burning', 'common', 'abnormal', 'uncontrolled'

Drugs Side Effects

The metadata comprises comprehensive details regarding a wide range of pharmaceuticals used to treat conditions ranging from acne to cancer and heart disease. It also provides an insight into their possible side effects. These details are not limited to the generic name of the drug but also

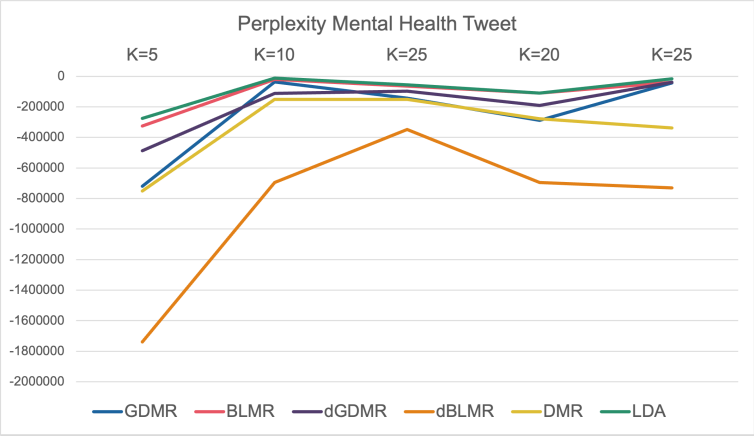


Figure 5.6: Perplexity comparison for Mental Health Tweet dataset

Table 5.6: Time complexity comparison for different model at varying topic levels (K) on the Symptom for disease (cancer) dataset. (min)

K	5	10	15	20	25
LDA	13.67	17.771	19.138	20.505	21.5986
DMR	10.12	13.156	14.168	15.18	15.9896
GDMR	8.96	11.648	12.544	13.44	14.1568
BLMR	8.16	10.608	11.424	12.24	12.8928
dBDMR	27.19	35.347	38.066	40.785	42.9602
DGMR	28.17	30.987	39.438	45.072	47.889

extend to its drug class, the different brand names it may be sold under, its activity level, whether it is a prescription drug or not, its categorization concerning pregnancy safety, its schedule under the Controlled Substances Act, possible interactions with alcohol, and its user ratings.

Our analysis was focused primarily on the features pertaining to medical conditions, and we employed topic modeling to discern patterns and topics based on the side effects linked to various drugs. We have provided the outcomes of our topic modeling for a few conditions, such as hypertension, in Tables 5.7,5.8 and Fig. 5.9 and 5.10. Fig. 5.10 illustrates that the dGDMR model has less perplexity compared to other models.

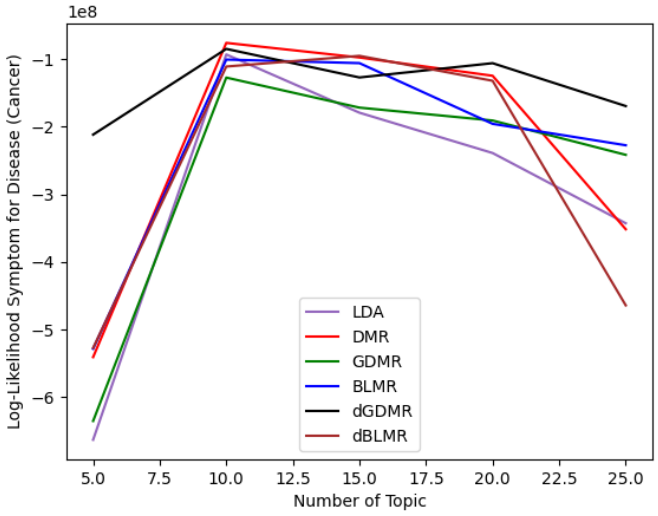


Figure 5.7: Log-likelihood for Symptom for Disease (Cancer)

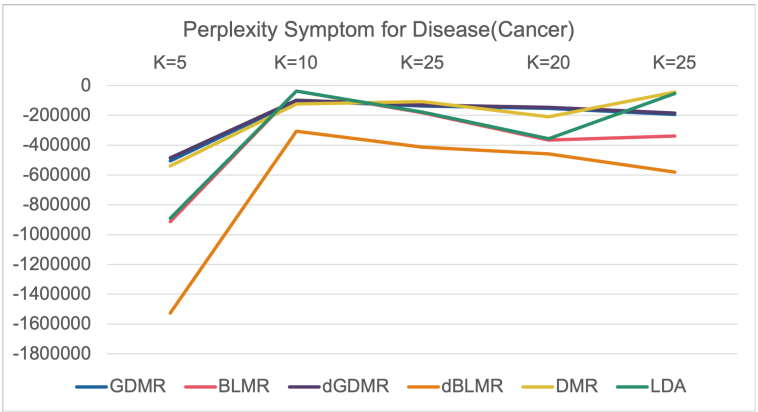


Figure 5.8: Perplexity comparison for Symptom for Disease (Cancer)

Table 5.7: Common topics identified with BLMPCA model in the Drugs Side Effects (hypertension) dataset, each defined by a set of keywords

Topic No	Topics
Topic1	'concentrate', 'chest', 'balance', 'trouble', 'hypertensioni', 'concentrate', 'issues', 'started', 'dizziness', 'challenging'
Topic2	'dizziness', 'hypertensioni', 'chest', 'headache', 'challenging', 'trouble', 'focus', 'experiencing', 'feeling', 'morning'
Topic3	'chest', 'headache', 'balance', 'dizzy', 'concentrate', 'issues', 'feeling', 'dizziness', 'having', 'developed'
Topic4	'focus', 'headache', 'issues', 'hypertensionive', 'experiencing', 'focusing', 'balance', 'feeling', 'challenging', 'experienced'
Topic5	'dizziness', 'focus', 'balance', 'trouble', 'chest', 'hypertensionalong', 'focusing', 'symptom', 'experiencing', 'morning'

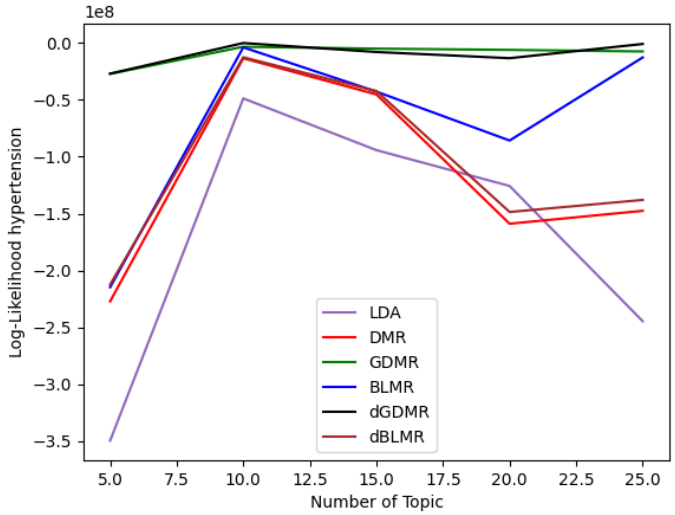


Figure 5.9: Log-likelihood for Drugs Side Effects (Hypertension)

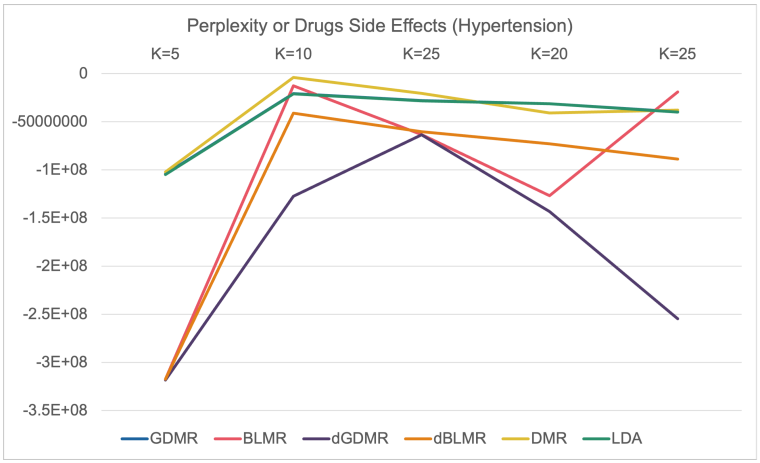


Figure 5.10: Perplexity comparison for Drugs Side Effects (Hypertension)

Table 5.8: Time complexity comparison for MPCA, GDMPCA, and BLMPCA at varying topic levels (K) on Drugs Side Effects (hypertension) dataset (min)

K	5	10	15	20	25
LDA	14.2	18.46	19.88	21.3	22.436
DMR	12.5	16.25	17.5	18.75	19.75
GDMR	3.62	4.70	5.06	5.43	5.71
BLMR	2.22	2.88	3.10	3.33	3.50
dGDMR	12.25	13.47	17.15	19.65	20.82
dBLMR	10.57	13.74	14.79	15.85	16.70

Chapter 6

Integration of Neural Embeddings and Probabilistic Models in Topic Modeling

6.1 Introduction

The pervasive inundation of digital textual data in today's information era underscores the imperative need for robust analytical methodologies. As diverse sectors, from academia to industry, grapple with vast swathes of unstructured text, the quest for tools capable of deciphering, organizing, and extracting meaningful insights has gained paramount importance. At the intersection of this quest lies the domain of document clustering and topic modeling, which have emerged as pivotal techniques in the arsenal of textual data analytics [1].

In the expansive domain of text, document clustering and topic modeling stand out as two profoundly researched problems due to their multifarious applications. Document clustering, which endeavors to aggregate similar documents into cohesive groups, serves as an indispensable tool for document organization, browsing, summarization [6], classification [5], and retrieval [151]. Concurrently, topic modeling employs probabilistic generative models to unearth the latent semantics permeating a collection of documents, a technique that has garnered significant accolades for its prowess in text analysis [8]. Among the suite of tools designed for revealing common themes and narratives in text, topic models, especially prominent ones like Latent Dirichlet Allocation (LDA) [13] and Non-Negative Matrix Factorization (NMF) [72], have been particularly influential. These

conventional models characterize documents as a ”bag of words” [120], envisioning each as a composite mixture of latent topics. However, this paradigm is not without its caveats.

A notable drawback of these models arises from their reliance on the bag-of-words framework. This representation simplifies texts to mere word counts, often sidelining the nuanced semantic connections that exist between words. Consequently, because such models do not factor in the contextual positioning and interplay of words within a sentence, there's a palpable risk that the bag-of-words input might not capture the true essence and intricacies of the document's content [152].

In response to the limitations posed by traditional models, the field of natural language processing has witnessed a surge in the adoption of text embedding technique s[153]. At the forefront of this transformative shift is the Bidirectional Encoder Representations from Transformers (BERT) framework and its assorted derivatives [23]. These methodologies have garnered significant attention and acclaim for their capability to generate contextually enriched vector representations of both words and sentences. Unlike their predecessors, BERT and its variations delve deeper by considering the positional interplay of words, thereby yielding embeddings that capture the nuanced semantics and intricacies inherent within a language structure [154, 23].

Embedding techniques, with their advanced capabilities, have permeated a wide array of tasks in the domain of natural language processing [155]. These tasks span from classification endeavors to the powering of sophisticated neural search engines. Given the enhanced contextual representations these techniques afford, their application in topic modeling has drawn increasing attention from the research community [156]. An instance of the effectiveness of clustering embeddings can be seen in the use of centroid-based techniques. Their findings indicated that these techniques provide a compelling substitute for conventional methods such as LDA in terms of efficient topic representation. The authors of the study employed a methodology that involved creating topic representations by embedding words and then identifying those words that were located near the centroid of a cluster [25].

Top2Vec, a method that utilizes Doc2Vec, can be compared to the current topic. Top2Vec excels at simultaneously embedding topics, documents, and word vectors. Following the methodology proposed by [25], Top2Vec utilizes document clustering to identify the central theme of a topic by

identifying words that are closely associated with the centroid of a cluster. A captivating aspect of this methodology is its clustering strategy [26, 157].

The intricate domains of document clustering and topic modeling share a symbiotic relationship, each enriching the other's capabilities. When delving into the mechanisms of topic models, one observes their adeptness at discerning the subtle, underlying semantics woven into a collection of documents. This distilled semantic essence offers a more refined lens for delineating distinct document groups compared to relying solely on crude term features. Traditional document clustering techniques, for the most part, employ the bag-of-words (BOW) model. While this approach is straightforward, basing representations purely on raw terms occasionally leaves them wanting in terms of capturing the entirety of semantic nuances.

Topic modeling, on the other hand, exhibits a more sophisticated approach. It congregates words bearing semantic resemblance under collective themes or "topics". An inherent advantage of this method is its ability to perceive synonymous words as conceptually identical, obviating the limitations of exact term matching. Furthermore, by transforming the document corpus into a topic-centric space, topic models effectively filter out extraneous noise associated with similarity measurements. As a result, the innate groupings within a corpus emerge more distinctly, paving the way for more precise and coherent clustering [4].

Topic modeling benefits from reciprocal document clustering. It distills both localized topics specific to document clusters and global topics that span multiple clusters. Google Scholar, for example, contains academic papers from math, biology, computer science, and economics. Each discipline or "group" has its own topical themes in this vast expanse. Computer science literature may cover machine learning, operating systems, and networking, while economics literature may cover financial economics, entrepreneurial theories, or mathematical economics [27, 10].

Despite these domain-specific topics, "global topics" exist that transcend disciplines. Academic papers may include literature reviews, experimental results, and funding and support acknowledgements. This scenario benefits from clustering, which reveals these document groups. After uncovering these latent groupings, one can identify topics that are unique to each group and those that are common to all. Granular topic demarcation opens many doors. Localized topics can be concise summaries or browsing aids for document groups. In contrast, global topics can help identify

common or background words, providing a complete document collection view [27, 28, 158].

The limitations of conventional topic modeling methods lie in their inability to capture the intricate dynamics within document groups. The current iterations of such models often lack the capacity to encapsulate the evolving dynamics within document collections. Consequently, they tend to yield generalized topics that obscure both local and global thematic nuances, thereby constraining comprehension and practical utility [13, 12, 16, 34].

A feasible strategy involves decoupling these two endeavors. Initially, topic models can be employed to map documents into thematic domains, followed by the application of K-means clustering to delineate clusters [29]. Subsequently, conventional clustering techniques can identify clusters, while topic modeling techniques can extract localized topics pertinent to each cluster and overarching topics unrelated to any specific cluster. Integration of cluster labels into the model's framework facilitates enhanced model performance [3, 4].

Nevertheless, a simplistic segregation of these tasks risks oversimplifying the intricate interplay between document clustering and topic modeling. A more sophisticated approach acknowledges the symbiotic relationship between these processes. Enhanced clustering methodologies can enrich the sophistication of topic models, and conversely, refined topic modeling techniques can inform more precise clustering outcomes. Treating these tasks in isolation may curtail their synergistic potential and overall effectiveness [29].

The Bert-Topic embedding stands as a testament to the advancements in natural language processing, employing the state-of-the-art capabilities of Bidirectional Encoder Representations from Transformers (BERT) to generate rich contextual word and sentence vector representations [23, 24]. It captures semantic intricacies, allowing for deeper topic discernment and superior document representation. On a parallel front, the Multi-Grain Clustering Topic Model (MGCTM) has been a beacon in topic modeling, offering structured granularity to elucidate both micro- and macro-level topics in text data [27]. By incorporating Generalized Dirichlet and Beta-Liouville distributions alongside Bert-Topic embedding into the MGCTM framework, we enhance its modeling versatility, extending the capabilities of the original model.

In this chapter, we provide our proposed models, namely the multi-grain Generalized Dirichlet Bert-Topic model (MGGDBTM) and the multi-grain Beta-Liouville Bert-Topic model (MGBLBTM). The initial step is delineating the characteristics of the fitting distribution associated with each model. Following this, we utilize variational approaches in order to ascertain the parameters of the distribution. This section is concluded by providing a comprehensive account of the learning mechanism for these models.

6.2 Multi-grain Generalized Dirichlet Bert-topic Model

The generalized Dirichlet (GD) distribution, introduced in [90], offers a more expansive covariance structure than the Dirichlet distribution and overcomes its constraints, including assumptions of negative correlation and equal confidence [159]. As such, it stands out as a preferred prior in Bayesian learning, as highlighted by [91, 143]. Moreover, [91] accentuates its viability as an alternative for the Dirichlet distribution when clustering count data with mixture models, given its role as a conjugate prior for the multinomial distribution and its robust covariance matrix. Building on the GD foundation, the multi-grain clustering topic model (MGCTM) [27] will be enhanced using the variational Bayes method. Given insights from [92], it's anticipated that the MGGDCTM model will surpass the performance of Dirichlet-based models, especially since the GD encompasses the Dirichlet distribution as a specific subset.

In the MGGDCTM model, we posit a corpus comprised of N documents, represented as $d \in \{1, 2, \ldots, N\}$. These documents are inherently categorized into J groups, denoted as $j \in \{1, 2, \ldots, J\}$. Each group, j, is characterized by K local topics specific to that group, symbolized as $\beta_j^{(l)}$. These local topics capture the unique semantics of each group. Additionally, every group j is associated with its distinct local GD prior, $\xi_j^{(l)}$. For documents within group j, their local topic proportion vectors are drawn from $\xi_j^{(l)}$. Beyond the group-specific local topics, we posit the existence of a single collection of R global topics, $\beta_j^{(g)}$, that are common across all groups. These global topics represent the overarching semantics present throughout the entire dataset. All documents share a universal GD prior, $\xi^{(g)}$, which determines the proportion vectors for these global topics. Additionally, a global multinomial prior, π, determines the group affiliation of each document.

Documents are linked to a specific group and have distributions related to both local and global topics. We utilize a Bernoulli variable, sourced from a Beta prior, to determine if a word originates from local or global topics. To produce a document, we select a group and then use various priors and distributions to determine the word's source, either local or global topics. The choice of topic source dictates the distribution used to produce each word. The generative process of a document in MGGDCTM model can be summarized as follows:

(1) $\nu \sim Multinomial(\pi)$

(2) Local Topic $m_{\nu}^{l} \sim GD(\xi_{\nu}^{l})$

(3) Global Topic $m^{g} \sim GD(\xi^{g})$

(4) For each word w:

(a) Binary indicator $\delta \sim bernouli(w)$:

i. If $\delta = 1$

A. Local topic $z_{\nu}^{l} \sim Multinomial(m_{\nu}^{l})$

B. $w \sim Multinomial(\Omega_{l})$

i. If $\delta = 0$

A. Global topic $z^{g} \sim Multinomial(m_{g})$

B. $w \sim Multinomial(\Omega_{g})$

In the process of applying a topic model prior to clustering, it is customary to assign predetermined values to the latent variables of the topic model in the MGGDCTM framework, followed by the optimization of the mixture model. When the clustering process is performed as the initial step, the latent variables of the mixture model are established, allowing us to subsequently concentrate on optimizing the topic model. However, when both components are conducted in conjunction, the objective is to concurrently maximize the latent variables of both.

The Generalized Dirichlet distribution, denoted as $GD(\xi)$, functions in a space of d dimensions. This distribution has parameters symbolized by ξ. These parameters can be comprehensively expressed as a sequence:

$(a_1, b_1, \ldots, a_d, b_d)$. When discussing the distribution, we refer to its probability distribution function using the notation p. One of the significant variables in this context is Λ_i. This variable is deduced by computing the difference between certain parameters: specifically, b_i, a_{i+1}, and b_{i+1}. Mathematically, the relationship can be captured by the equation: $\Lambda_i = b_i - a_{i+1} - b_{i+1}$ [46].

$$p(m_1, \ldots, m_d|\xi) = \prod_{i=1}^{d} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} m_i^{a_l - 1}(1 - \sum_{j=1}^{i} m_j)^{\Lambda_i} \tag{198}$$

Having established our Generalized Dirichlet (GD) prior, our next course of action is to construct the word-topic probability matrix, which we label as Ω. By operating under the presumption of conditional independence among the variables, it becomes feasible to deduce the joint distribution. The mathematical representation for this distribution is:

$$p(m, z, w, |\xi, \Omega) = p(m|\xi)p(w|z, \Omega)p(z|m) \tag{199}$$

In this formulation, the symbol z corresponds to the collection of underlying topics. The strength of this representation lies in its ability to factor in different variables and their interrelations.

The process of integrating across the m parameters and the topic space yields the following outcome:

$$\begin{aligned} p(w|\xi, \Omega) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \int m^{a_i - 1}(1 - \sum_{j=1}^{i} m_j)^{\Lambda_i} \\ \times \prod_{n=1}^{N} \prod_{j=1}^{d+1} \prod_{j=1}^{V} (m\Omega_{ij})^{w_n^j} d\theta \end{aligned} \tag{200}$$

In Equation 200, the variables ξ and Ω represent the parameters at the corpus level, whilst the variable m represents the parameter at the document level. On the other hand, the variables z and w represent the parameters at the word level.

6.2.1 Variational Inference

In the MGGDCTM model, accurately estimating the posterior distribution $P(\nu, w, m^{(l)}, m^{(g)}, \delta, z^{(l)}, z^{(g)}|w, m)$ of the latent variables $\boldsymbol{B} = \{\nu, w, m^{(l)}$ $, m^{(g)}, \delta, z^{(l)}, z^{(g)}\}$ is challenging. This estimation depends on observed data and the model parameters $\boldsymbol{\Theta} = \{\pi, \lambda, \xi^{(l)}, \xi^{(g)}, \beta^{(l)}, \beta^{(g)}\}$.

$$P(\nu, w, m^{(l)}, m^{(g)}, \delta, z^{(l)}, z^{(g)}|w, m) = \frac{P(\nu, w, \delta, z^{(l)}, z^{(g)}|m^{(l)}, m^{(g)})}{P(w|m^{(l)}, m^{(g)}))} \tag{201}$$

Given these premises, the joint distribution of the posterior can be expressed as follows:

$$\begin{aligned} p(W, \Theta) &= p(W|z^{(l)}, \Omega^{(l)}, \theta^{(l)}, \sigma, \tau, y, z^{(g)}, \Omega^{(g)}, \theta^{(g)}, \alpha, \beta) \\ &= p(W|z^{(l)}, \Omega^{(l)}, \delta, \Omega^{(g)}, z^{(g)}) p(z^{(l)}|\theta^{(l)}) p(\theta^{(l)}|\sigma, \tau, y) p(y|\pi) p(\sigma|\nu, \vartheta) \\ &\quad \times p(\tau|s, t) p(\theta^{(l)}|\iota, \varrho) p(z^{(g)}|\theta^{(g)}) p(\theta^{(g)}|\beta, \alpha) p(\alpha|g, h) p(\beta|a, b) \\ &\quad p(\delta|\omega) p(\omega|\gamma) p(\Omega^{(l)}|\lambda) p(\Omega^{(g)}|\kappa) p(\theta^{(g)}|\chi, s) \end{aligned} \tag{202}$$

Eq. 202 describes the joint probability distribution of the observed words (W) and the set of all parameters in the model (Θ). The observed words are represented as W, while Θ encompasses all model parameters. The local topic assignments for words in documents are denoted by $z^{(l)}$, and the local word-topic probability matrix, which represents the probability of each word given a local topic, is denoted by $\Omega^{(l)}$. The local topic proportions for each document are represented by $\theta^{(l)}$, and these proportions are governed by the parameters σ and τ. The document group assignments are denoted by y.

The global topic assignments for words in documents are represented by $z^{(g)}$, and the global word-topic probability matrix, which represents the probability of each word given a global topic, is denoted by $\Omega^{(g)}$. The global topic proportions for each document are denoted by $\theta^{(g)}$, with α and β as the parameters for the global topic proportions distribution. The prior probability vector for the group assignments is represented by π, and the hyperparameters for the prior distribution of σ are ν and ϑ. Similarly, s and t are the hyperparameters for the prior distribution of τ.

The local topic proportions are influenced by the hyperparameters ι and ϱ. The hyperparameters for the prior distribution of α are g and h, while a and b are the hyperparameters for the prior distribution of β. The binary indicator for local or global topic generation is represented by δ, with ω as the hyperparameters for its prior distribution. The hyperparameters for the prior distribution of ω are represented by γ. The prior distribution of $\Omega^{(l)}$ is governed by the hyperparameters λ, and the prior distribution of $\Omega^{(g)}$ is governed by κ. Lastly, the global topic proportions are influenced by the hyperparameters χ and s.

Gibbs sampling is typically used in traditional topic models due to a particular conjugate relationship. However, the estimation of the posterior distribution becomes complex due to a mismatch in the prior topic distribution for phrases. To tackle this, variational inference is utilized for approximating the distribution. The main objective is to minimize the Kullback-Leibler (KL) divergence between the actual and variational posteriors, as indicated in various studies [94, 128, 27]. This issue is approached as an optimization problem. Research suggests that enhancing the variational posterior probability and reducing the KL divergence can lead to a better evidence lower bound (ELBO). Ultimately, variational inference aims to align variational distributions closely with the true posterior [66, 129, 37, 128].

Variational inference is chosen over traditional Bayesian methods like Gibbs sampling [160] because, despite potentially providing more accurate parameter estimates, these algorithms can require an extended time to converge. In contrast, variational methods introduce a distribution $Q(\Theta)$ as an approximate representation of $P(W|\Theta)$, the sought-after posterior distribution. This strategy effectively addresses the limitations of classic Bayesian methods by estimating, rather than precisely calculating, the posterior. Our methodology involves assessing the closeness of the posterior and variational distributions through the Kullback-Leibler (KL) divergence. A KL divergence of 0 indicates similarity between two distributions. The KL divergence between $Q(\Theta)$ and $P(W|\Theta)$ is defined as,

$$KL(Q||P) = -\int Q(\Theta) \ln \left(\frac{p(W|\Theta)}{Q(\Theta)}\right) d\Theta \tag{203}$$

Simplifying this equation yields:

$$KL(Q||P) = \ln p(W) - \mathcal{L}(Q) \tag{204}$$

where,

$$\mathcal{L}(Q) = \int Q(\Theta) \ln \left(\frac{p(W,\Theta)}{Q(\Theta)}\right) d\Theta \tag{205}$$

The principle underlying Eqs. 203, 204 and 205 is that maximizing the lower bound $\mathcal{L}(Q)$ effectively reduces the KL divergence towards 0. Given the intractability of the true posterior, mean-field theory [161] is introduced, assuming the parameters are independent and identically distributed. From this standpoint, the distribution of variational parameters can be represented as a product of individual parameters, expressed as $Q(\Theta) = \prod_{j=1}^{J} \Theta_j$, where J denotes the total number of parameters. The optimal solution for each parameter is then determined by the subsequent equation,

$$Q_j(\Theta_j) = \frac{\exp(\ln p(W,\Theta))_{\neq j}}{\int \exp(\ln p(W,\Theta))_{\neq j} d\Theta} \tag{206}$$

Based on the given equation, it's clear that the optimal solution for the parameter Θ_j is determined by computing the expectations with respect to all parameters except Θ_j. This necessitates an appropriate initialization at the beginning of the algorithm. Subsequently, the variational solutions for each parameter are iteratively updated, which contributes to the maximization of the lower bound. Upon reaching convergence, the algorithm yields the optimal solution for all the parameters in the model.

6.2.2 Variational solutions for MGGDCTM

All parameters used in the equations throughout this paper are defined in Table 6.1.

Determining the variational solutions for Eq 202, the subsequent equations are derived:

$$Q(y) = \prod_{d=1}^{D} \prod_{l=1}^{L} r_{dl}^{y_{dl}} \tag{207}$$

Table 6.1: Definitions of Parameters Used in the Equations MGGDCTM model

D	Number of documents
L	Number of document groups
N_d	Number of words in document d
K	Number of topics
V	Vocabulary size
r_{dl}	Document-group assignment probability for document d and group l
y_{dl}	Indicator variable if document d is assigned to group l
$\phi_{dnk}^{(g)}$	Global topic assignment probability for word n in document d and topic k
$z_{dnk}^{(g)}$	Indicator variable if word n in document d is assigned to global topic k
$\phi_{dnk}^{(l)}$	Local topic assignment probability for word n in document d and topic k
$z_{dnk}^{(l)}$	Indicator variable if word n in document d is assigned to local topic k
σ_{lk}, τ_{lk}	Parameters for the local topic proportions distribution
$\vartheta_{lk}^*, \nu_{lk}^*$	Hyperparameters for the prior distribution of σ_{lk}
t_{lk}^*, s_{lk}^*	Hyperparameters for the prior distribution of τ_{lk}
$\Omega_{kv}^{(l)}$	Local word-topic probability for word v given topic k
λ_{kv}^*	Hyperparameters for the prior distribution of $\Omega_{kv}^{(l)}$
α_{lk}, β_{lk}	Parameters for the global topic proportions distribution
h_{lk}^*, g_{lk}^*	Hyperparameters for the prior distribution of α_{lk}
a_{lk}^*, b_{lk}^*	Hyperparameters for the prior distribution of β_{lk}
$\Omega_{kv}^{(g)}$	Global word-topic probability for word v given topic k
κ_{kv}^*	Hyperparameters for the prior distribution of $\Omega_{kv}^{(g)}$
$\theta_{dk}^{(l)}$	Local topic proportion for document d and topic k
$\iota_{dk}^*, \varrho_{dk}^*$	Hyperparameters for the local topic proportions
$\theta_{dk}^{(g)}$	Global topic proportion for document d and topic k
χ_{dk}^*, s_{dk}^*	Hyperparameters for the global topic proportions
$\rho_{dl}, \delta_{dnk}^{(l)}, \delta_{dnk}^{(g)}$	Intermediate variables for probability calculations
$\mathcal{R}^{(l)}, \mathcal{R}^{(g)}$	Taylor series approximations for the digamma and trigamma functions

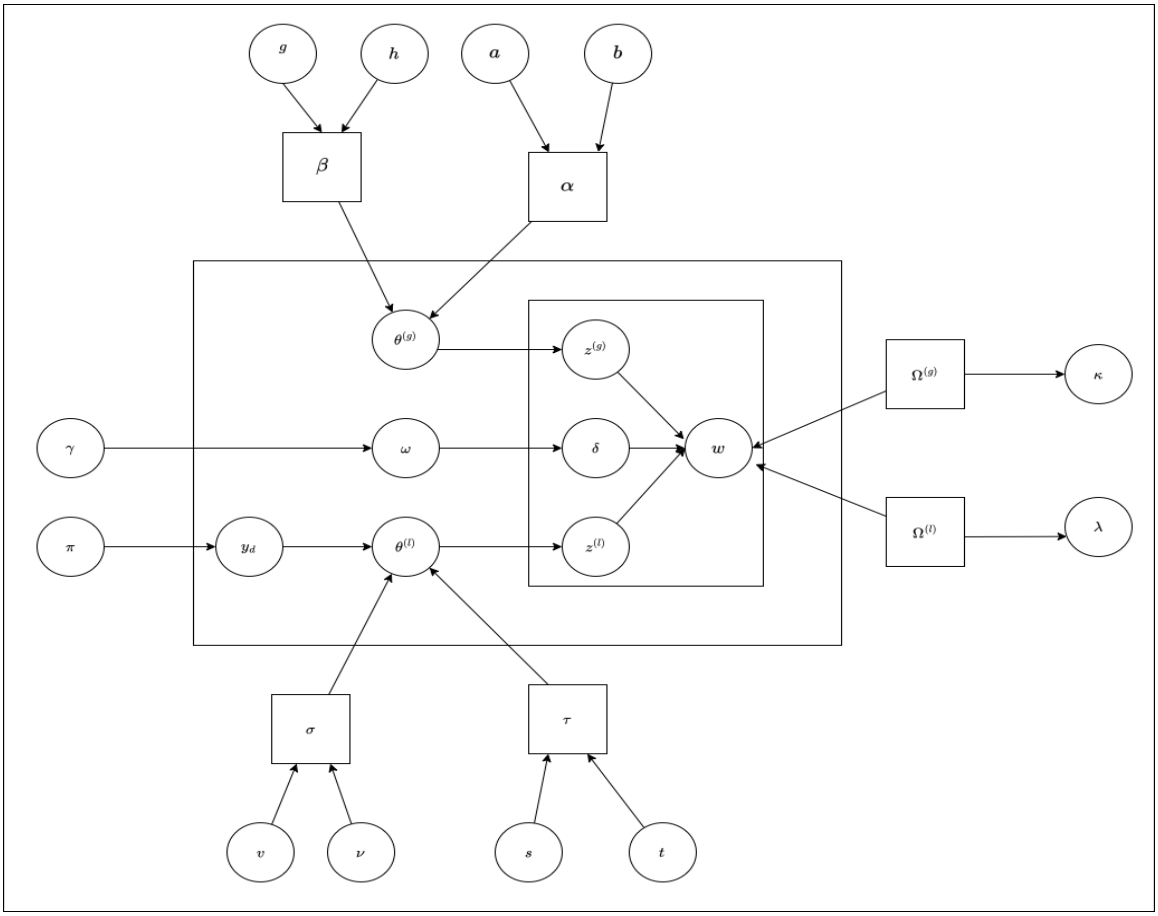


Figure 6.1: Graphical representation of MGGDCTM

$$Q(z^{(g)}) = \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{K=1}^{K} \phi_{dnk}^{(g) z_{dnk}^{(g)}}, \quad Q(z^{(l)}) = \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{K=1}^{K} \phi_{dnk}^{(l) z_{dnk}^{(l)}} \tag{208}$$

$$Q(\sigma) = \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{{\vartheta^*}_{lk}^{\nu_{lk}^*}}{\Gamma(\nu_{lk}^*)} \sigma_{lk}^{\nu_{lk}^* - 1} e^{-\vartheta_{lk}^* \sigma_{lk}}, \quad Q(\tau) = \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{{t^*}_{lk}^{s_{lk}^*}}{\Gamma(s_{lk}^*)} \tau_{lk}^{s_{lk}^* - 1} e^{-t_{lk}^* \tau_{lk}} \tag{209}$$

$$Q(\Omega^{(l)}) = \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma\left(\sum_{v=1}^{V} \lambda_{kv}^*\right)}{\prod_{v=1}^{V} \Gamma(\lambda_{kv}^*)} \Omega_{kv}^{(l) \lambda_{kv}^* - 1}, \quad Q(\Omega^{(g)}) = \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma\left(\sum_{v=1}^{V} \kappa_{kv}^*\right)}{\prod_{v=1}^{V} \Gamma(\kappa_{kv}^*)} \Omega_{kv}^{(g) \kappa_{kv}^* - 1} \tag{210}$$

$$Q(\alpha) = \prod_{l=1}^{L}\prod_{k=1}^{K} \frac{{h_{lk}^*}^{s_{lk}^*}}{\Gamma(g_{lk}^*)} \alpha_{lk}^{g_{lk}^*-1} e^{-h_{lk}^*\alpha_{lk}}, \quad Q(\beta) = \prod_{l=1}^{L}\prod_{k=1}^{K} \frac{{a_{lk}^*}^{b_{lk}^*}}{\Gamma(b_{lk}^*)} \beta_{lk}^{b_{lk}^*-1} e^{-a_{lk}^*\beta_{lk}} \tag{211}$$

$$Q(\theta^{(l)}) = \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(\iota_{dk}^* + \varrho_{dk}^*)}{\Gamma(\iota_{dk}^*)\Gamma(\varrho_{dk}^*)} \theta_{dk}^{(l)\iota_{dk}^*-1} \left(1 - \sum_{j=1}^{k} \theta_{dj}^{(l)}\right)^{\zeta_{dk}^*} \tag{212}$$

$$Q(\theta^{(g)}) = \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(\chi_{dk}^* + s_{dk}^*)}{\Gamma(\chi_{dk}^*)\Gamma(s_{dk}^*)} \theta_{dk}^{(g)\chi_{dk}^*-1} \left(1 - \sum_{j=1}^{k} \theta_{dj}^{(g)}\right)^{\zeta_{dk}^*} \tag{213}$$

where,

$$\begin{aligned} r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L}\rho_{dl}}, \quad \phi_{dnk}^{(l)} = \frac{\delta_{dnk}^{(l)}}{\sum_{k=1}^{K}\delta_{dnk}^{(l)}}, \quad \pi_l = \frac{1}{D}\sum_{d=1}^{D} r_{dl}, \\ \phi_{dnk}^{(g)} = \frac{\delta_{dnk}^{(g)}}{\sum_{k=1}^{K}\delta_{dnk}^{(g)}}. \end{aligned} \tag{214}$$

$$\rho_{dl} = \exp\left\{\ln \pi_l + R_l + \sum_{k=1}^{K}(\sigma_{lk}-1)(\ln \theta_{dk}^{(l)}) + \gamma_{lk}(\langle\rangle 1 - \sum_{j=1}^{k}\theta_{dj}^{(l)})\right\} \tag{215}$$

$$\delta_{dnk}^{(l)} = \exp(\ln \Omega_{kv}^{(l)} + \ln \theta_{dk}^{(l)}), \quad \delta_{dnk}^{(g)} = \exp(\ln \Omega_{kv}^{(g)} + \ln \theta_{dk}^{(g)}) \tag{216}$$

In this context, $\mathcal{R}^{(l)}$ and $\mathcal{R}^{(g)}$ represent the Taylor series approximations of $\ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)+\Gamma(\tau)}$ and $\ln \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}$, respectively, which are expressed as follows:

$$
\begin{aligned}
R^{(l)} = & \ln \frac{\Gamma(\overline{\sigma}+\overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\left[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\sigma})\right]\left(\langle \ln \sigma \rangle - \ln \overline{\sigma}\right) \\
& + \overline{\tau}\left[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\tau})\right]\left(\langle \ln \tau \rangle - \ln \overline{\tau}\right) \\
& + 0.5\overline{\sigma}^2\left[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\sigma})\right]\langle(\ln \sigma - \ln \overline{\sigma})^2\rangle \\
& + 0.5\overline{\tau}^2\left[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\tau})\right]\left(\langle \ln \tau - \ln \overline{\tau})^2\rangle\right. \\
& + \overline{\sigma}\overline{\tau}\Psi'(\overline{\sigma}+\overline{\tau})\left(\langle \ln \sigma \rangle - \ln \overline{\sigma}\right)\left(\langle \ln \tau \rangle - \ln \overline{\tau}\right)
\end{aligned} \tag{217}
$$

$$
\begin{aligned}
R^{(g)} = & \ln \frac{\Gamma(\overline{\alpha}+\overline{\beta})}{\Gamma(\overline{\alpha})\Gamma(\overline{\beta})} + \overline{\alpha}\left[\Psi(\overline{\alpha}+\overline{\beta}) - \Psi(\overline{\alpha})\right]\left(\langle \ln \alpha \rangle - \ln \overline{\alpha}\right) \\
& + \overline{\beta}\left[\Psi(\overline{\alpha}+\overline{\beta}) - \Psi(\overline{\beta})\right]\left(\langle \ln \beta \rangle - \ln \overline{\beta}\right) \\
& + 0.5\overline{\alpha}^2\left[\Psi'(\overline{\alpha}+\overline{\beta}) - \Psi'(\overline{\alpha})\right]\langle(\ln \alpha - \ln \overline{\alpha})^2\rangle \\
& + 0.5\overline{\beta}^2\left[\Psi'(\overline{\alpha}+\overline{\beta}) - \Psi'(\overline{\beta})\right]\left(\langle \ln \beta - \ln \overline{\beta}\right)^2\rangle \\
& + \overline{\alpha}\overline{\beta}\Psi'(\overline{\alpha}+\overline{\beta})\left(\langle \ln \alpha \rangle - \ln \overline{\alpha}\right)\left(\langle \ln \beta \rangle - \ln \overline{\beta}\right)
\end{aligned} \tag{218}
$$

$$
\begin{aligned}
v^*_{ik} = & \, v_{lk} + \sum_{d=1}^{D}\langle y_{dl}\rangle[\Psi(\overline{\sigma}_{lk}+\overline{\tau}_{lk}) - \Psi(\overline{\sigma}_{lk}) \\
& + \overline{\tau}_{lk}\Psi'(\overline{\sigma}_{lk}+\overline{\tau}_{lk})\left(\langle \ln \tau_{lk}\rangle - \ln \overline{\tau}_{lk}\right)\overline{\sigma}_{lk}
\end{aligned} \tag{219}
$$

$$
\begin{aligned}
s^*_{lk} = & \, s_{lk} + \sum_{d=1}^{D}\langle y_{dl}\rangle[\Psi(\overline{\tau}_{lk}+\overline{\sigma}_{lk}) - \Psi(\overline{\tau}_{lk}) \\
& + \overline{\sigma}_{lk}\Psi'(\overline{\tau}_{lk}+\overline{\sigma}_{lk})\left(\langle \ln \sigma_{lk}\rangle - \ln \overline{\sigma}_{lk})\right)\overline{\tau}_{lk}
\end{aligned} \tag{220}
$$

$$
v^*_{lk} = v_{lk} - \sum_{d=1}^{D} y_{dl}\ \ln \theta_{dk}, \quad t^*_{lk} = t_{lk} - \sum_{d=1}^{D}\langle y_{dl}\rangle\langle \ln\left(1 - \sum_{j=1}^{K}\theta_{dj}\right)\rangle \tag{221}
$$

$$\iota_{dk}^{*} = \iota_{dk} + \sum_{n=1}^{N_d} \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \sigma_{lk}, \quad \varrho_{dk}^{*} = \varrho_{dk} + \sum_{l=1}^{L} \langle y_{dl} \rangle \tau_{lk} + \sum_{kk=k+1}^{K} \phi_{dn(kk)} \tag{222}$$

$$\lambda_{kv}^{*} = \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} \phi_{dnk} w_{dnv}, \quad \pi_l = \frac{1}{D} \sum_{d=1}^{D} r_{dl} \tag{223}$$

$$\begin{aligned} a_{ik}^{*} = a_{lk} + \sum_{d=1}^{D} [& \Psi(\overline{\alpha}_{lk} + \overline{\tau}_{lk}) - \Psi(\overline{\alpha}_{lk}) \\ & + \overline{\beta}_{lk} \Psi'(\overline{\alpha}_{lk} + \overline{\beta}_{lk}) \left(\ln \beta_{lk} - \ln \overline{\beta}_{lk} \right) \overline{\alpha}_{lk} \end{aligned} \tag{224}$$

$$\begin{aligned} g_{lk}^{*} = g_{lk} + \sum_{d=1}^{D} [& \Psi(\overline{\beta}_{lk} + \overline{\alpha}_{lk}) - \Psi(\overline{\beta}_{lk}) \\ & + \overline{\alpha}_{lk} \Psi'(\overline{\beta}_{lk} + \overline{\alpha}_{lk}) \left(\ln \alpha_{lk} - \ln \overline{\sigma}_{lk} \right)) \overline{\beta}_{lk} \end{aligned} \tag{225}$$

$$b_{lk}^{*} = b_{lk} - \sum_{d=1}^{D} \ln \theta_{dk}^{(g)}, \quad h_{lk}^{*} = h_{lk} - \sum_{d=1}^{D} \ln \left(1 - \sum_{j=1}^{K} \theta_{dj}^{(g)} \right) \tag{226}$$

$$\chi_{dk}^{*} = \chi_{dk} + \sum_{n=1}^{N_d} z_{dnk}^{(g)} + \sum_{l=1}^{L} \alpha_{lk}, \quad s_{dk}^{*} = s_{dk} + \sum_{l=1}^{L} \beta_{lk} + \sum_{kk=k+1}^{K} \phi_{dn(kk)}^{(g)} \tag{227}$$

$$\kappa_{kv}^{*} = \kappa_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} \phi_{dnk}^{(g)} w_{dnv} \tag{228}$$

In the equations mentioned earlier, $(\bar{\cdot})$ symbolizes the average value of the variable. The estimations of these expectations and the mean, as referenced in [162], are presented as follows:

$$\ln \theta_{dk}^{(l)} = \sum_{j=1}^{k} \left(\Psi(\iota_{dk}) - \Psi(\iota_{dk} + \varrho_{dk}) \right) \tag{229}$$

$$1-\sum_{j=1}^{k}\theta_{dj}^{(l)}=\sum_{j=1}^{k}\left(\Psi(\varrho_{dk})-\Psi(\iota_{dk}+\varrho_{dk})\right) \tag{230}$$

$$\overline{\sigma}_{lk}=\frac{v_{lk}^{*}}{\vartheta_{lk}^{*}},\quad \ln\sigma_{uk}=\Psi(v_{lk}^{*})-\ln\vartheta_{lk}^{*} \tag{231}$$

$$(\ln\sigma_{lk}-\ln\overline{\sigma}_{lk})^{2}\ =[\Psi(v_{lk}^{*})-\ln v_{lk}^{*}]^{2}+\Psi'(v_{lk}^{*}) \tag{232}$$

$$\overline{\tau}_{lk}=\frac{s_{lk}^{*}}{t_{lk}^{*}},\quad \ln\tau_{lk}=\Psi(s_{lk}^{*})-\ln t_{lk}^{*} \tag{233}$$

$$(\ln\tau_{uk}-\ln\overline{\tau}_{uk})^{2}=[\Psi(s_{lk}^{*})-\ln s_{lk}^{*}]^{2}+\Psi'(s_{lk}^{*}) \tag{234}$$

$$z_{dnk}^{(l)}=\phi_{dnk},\quad y_{dl}=r_{dl},\quad \ln\Omega_{kv}^{(l)}=\Psi(\lambda_{kv})-\Psi\left(\sum_{f=1}^{V}\lambda_{kf}\right) \tag{235}$$

$$\theta_{dk}^{(g)}=\sum_{j=1}^{k}\left(\Psi(\chi_{dk})-\Psi(\chi_{dk}+s_{dk})\right) \tag{236}$$

$$1-\sum_{j=1}^{k}\theta_{dj}^{(g)}=\sum_{j=1}^{k}\left(\Psi(s_{dk})-\Psi(\chi_{dk}+s_{dk})\right) \tag{237}$$

$$\overline{\alpha}_{lk}=\frac{b_{lk}^{*}}{a_{lk}^{*}},\quad \ln\alpha_{lk}=\Psi(b_{lk}^{*})-\ln a_{lk}^{*} \tag{238}$$

$$(\ln\alpha_{lk}-\ln\overline{\alpha}_{lk})^{2}=[\Psi(b_{lk}^{*})-\ln b_{lk}^{*}]^{2}+\Psi'(b_{lk}^{*}) \tag{239}$$

$$\overline{\beta}_{lk}=\frac{g_{lk}^{*}}{h_{lk}^{*}},\quad \ln\beta_{lk}=\Psi(g_{lk}^{*})-\ln h_{lk}^{*} \tag{240}$$

$$\left(\ln \beta_{uk} - \ln \overline{\beta}_{uk}\right)^2 = \left[\Psi(g^*_{lk}) - \ln g^*_{lk}\right]^2 + \Psi'(g^*_{lk}) \tag{241}$$

$$z^{(g)}_{dnk} = \phi_{dnk}, \quad \ln \Omega^{(g)}_{kv} = \Psi(\kappa_{kv}) - \Psi\left(\sum_{f=1}^{V} \kappa_{kf}\right) \tag{242}$$

In the aforementioned equations, the symbols Ψ and Ψ' represent the digamma and trigamma functions, respectively.

The optimization of the model parameters was achieved by maximizing the lower bound, as demonstrated in the following:

$$\Omega^{(l)} = \sum_{d=1}^{D}\sum_{n=2}^{N_d}\sum_{v=1}^{V} \iota\varrho\theta^{(l)}_{dnk} w_{dnv}, \quad \Omega^{(g)} = \sum_{d=1}^{D}\sum_{n=2}^{N_d}\sum_{v=1}^{V} \chi s\theta^{(g)}_{dnk} w_{dnv} \tag{243}$$

Parameter Estimation

The terms in Eq. 202 that contain the GD parameters ξ are chosen:

$$\begin{aligned}\mathcal{L}[\xi] = &\sum_{m=1}^{M} (\log(\Gamma(\alpha_l + \beta_l)) - \log\Gamma(\alpha_l)) - \log(\Gamma(\beta_l))) \\ &+ \sum_{m=1}^{M} (\alpha_l(\Psi(\gamma_{ml} - \Psi(\gamma_{ml} + \delta_{ml})) + \beta_l(\Psi(\delta_{ml}) - \Psi(\delta_{ml} - \gamma_{ml})))\end{aligned} \tag{244}$$

The derivative of the above equation with respect to the GD parameters yields:

$$\frac{\partial \mathcal{L}[\xi]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M} (\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{245}$$

and

$$\frac{\partial \mathcal{L}[\xi]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M} (\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{246}$$

To solve the Newton-Raphson equation, acquiring the Hessian matrix within the parameter space is essential, as this matrix plays a crucial role in the optimization process:

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \alpha_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\alpha_l)] \tag{247}$$

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \beta_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\beta_l)] \tag{248}$$

$$\frac{\partial^2 \mathcal{L}[\xi]}{\partial \alpha_l \beta_l} = M[\Psi^{'}(\alpha_l + \beta_l)] \tag{249}$$

6.3 Multi-grain Beta-Liouville Bert-Topic Model

The MGBLBM model is built using the same principles as the MGGDBM model, with the key distinction lying in the prior used for topic proportions and its respective parameters. Specifically, the GD prior in the MGGDBM is substituted with the BL prior in the MGBLBM, resulting in an altered set of equations.

$$\begin{aligned} p(\theta^{(l)} \mid y, \mu, \sigma, \tau) =& \prod_{l=1}^{L} \prod_{k=1}^{K} \left(p(\theta_k^{(l)} \mid \mu_{lk}, \sigma_l, \tau_l) \right)^{y_{dl}} \\ =& \prod_{l=1}^{L} \prod_{k=1}^{K} \left[\frac{\Gamma\left(\sum_{k=1}^{K} \mu_{lk}\right)}{\prod_{k=1}^{K} \Gamma(\mu_{lk})} \frac{\Gamma(\sigma_l + \tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)} \left(\theta_{dk}^{(l)}\right)^{\mu_{lk}-1} \right. \\ & \left. \times \left(\sum_{k=1}^{K} \theta_{dk}^{(l)} \right)^{\sigma_l - \sum_{k=1}^{K} \mu_{lk}} \left(1 - \sum_{k=1}^{K} \theta_{dk}^{(l)} \right)^{\tau_l - 1} \right]^{y_{dl}} \end{aligned} \tag{250}$$

This implies that $\theta^{(l)}$ is considered a stochastic vector that follows a Beta-Liouville distribution with specified parameters $(\mu_{l1}, \mu_{l2}, ..., \mu_{lN_d}, \sigma_l, \tau_l)$. Expanding on this assumption, we can express the Gamma priors for these parameters as follows: $p(\mu_{lk}) = \mathcal{G}(\mu_{lk} \mid \upsilon_{lk}, \nu_{lk})$, $p(\sigma_l) = \mathcal{G}(\sigma_l \mid s_l, t_l)$, and $p(\tau_l) = \mathcal{G}(\tau_l \mid \Omega_l, \Lambda_l)$. These priors exhibit the same characteristics as those in the case of GD. Consequently, with these modifications, the variational distribution in Equation 250 will be replaced by the following:

$$p(\theta^{(l)}{}_d \mid c_d, R_d, Q_d) = \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{k=1}^{K} c_{dk}\right)}{\prod_{k=1}^{K} \Gamma(c_{dk})} \frac{\Gamma(R_d + Q_d)}{\Gamma(R_d)\Gamma(Q_d)} \left(\theta_{dk}^{(l)}\right)^{c_{dk}-1} \times \left[\sum_{k=1}^{K} \theta_{dk}^{(l)}\right]^{R_d - \sum_{k=1}^{K} c_{dk}} \left[1 - \sum_{k=1}^{K} \theta_{dk}^{(l)}\right]^{Q_d - 1} \tag{251}$$

By implementing these modifications, we are able to build the combined probability distribution of the posterior, assuming a BL prior for the topic proportions [163, 164], as follows:

$$
\begin{aligned}
p(W,\Theta) = & \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\left(\prod_{v=1}^{V}\left(\beta_{kv}^{(l)}\right)^{w_{dnv}}\right)^{z_{dnk}^{(l)}} \times \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\left(\theta_{dk}^{(l)}\right)^{z_{dnk}^{(l)}} \\
& \times \prod_{d=1}^{D}\prod_{l=1}^{L}\left[\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{k=1}^{K}\mu_{lk}\right)}{\prod_{k=1}^{K}\Gamma(\mu_{lk})}\frac{\Gamma(\sigma_l+\tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)}\left(\theta_{dk}^{(l)}\right)^{\mu_{lk}-1}\right. \\
& \left. \times \left(\sum_{k=1}^{K}\theta_{dk}^{(l)}\right)^{\sigma_l-\sum_{k=1}^{K}\mu_{lk}}\left(1-\sum_{k=1}^{K}\theta_{dk}^{(l)}\right)^{\tau_l-1}\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\left(\sum_{k=1}^{K}f_{dk}\right)}{\prod_{k=1}^{K}\Gamma(f_{dk})}\frac{\Gamma(g_d+h_d)}{\Gamma(g_d)\Gamma(h_d)}\left(\theta_{dk}^{(l)}\right)^{f_{dk}-1}\left(\sum_{k=1}^{K}\theta_{dk}^{(l)}\right)^{g_d-\sum_{k=1}^{K}f_{dk}} \\
& \times \left(1-\sum_{k=1}^{K}\theta_{dk}^{(l)}\right)^{h_d-1} \times \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right)}{\prod_{v=1}^{V}\Gamma(\lambda_{kv})}\left(\beta_{kv}^{(l)}\right)^{\lambda_{kv}-1} \\
& \times \prod_{l=1}^{L}\prod_{k=1}^{K}\frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})}\left(\mu_{lk}\right)^{\upsilon_{lk}-1}e^{-\nu_{lk}\mu_{lk}} \times \prod_{l=1}^{L}\frac{t_l^{s_l}}{\Gamma(s_l)}\left(\sigma_l\right)^{s_l-1}e^{-t_l\sigma_l} \\
& \times \prod_{l=1}^{L}\frac{\Lambda_l^{\Omega_l}}{\Gamma(\Omega_l)}\left(\tau_l\right)^{\Omega_l-1}e^{-\Lambda_l\tau_l} \\
& \times \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\left(\prod_{v=1}^{V}\left(\beta_{kv}^{(g)}\right)^{w_{dnv}}\right)^{z_{dnk}^{(g)}} \times \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\left(\theta_{dk}^{(g)}\right)^{z_{dnk}^{(g)}} \\
& \times \prod_{d=1}^{D}\prod_{l=1}^{L}\left[\prod_{k=1}^{K}\frac{\Gamma\left(\sum k=1^{K}B_{lk}\right)}{\prod_{k=1}^{K}\Gamma(B_{lk})}\frac{\Gamma(E_l+A_l)}{\Gamma(E_l)\Gamma(A_l)}\left(\theta_{dk}^{(g)}\right)^{B_{lk}-1}\right. \\
& \left. \times \left(\sum k=1^{K}\theta_{dk}^{(g)}\right)^{E_l-\sum k=1^{K}B_{lk}}\left(1-\sum k=1^{K}\theta_{dk}^{(g)}\right)^{A_l-1}\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\left(\sum k=1^{K}c_{dk}\right)}{\prod_{k=1}^{K}\Gamma(c_{dk})}\frac{\Gamma(R_d+Q_d)}{\Gamma(R_d)\Gamma(Q_d)}\left(\theta_{dk}^{(g)}\right)^{c_{dk}-1} \\
& \left(\sum_{k=1}^{K}\theta_{dk}^{(g)}\right)^{R_d-\sum k=1^{K}c_{dk}} \\
& \times \left(1-\sum k=1^{K}\theta_{dk}^{(g)}\right)^{Q_d-1} \times \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma\left(\sum v=1^{V}\rho_{kv}\right)}{\prod_{v=1}^{V}\Gamma(\rho_{kv})}\left(\beta_{kv}^{(g)}\right)^{\rho_{kv}-1} \\
& \times \prod_{l=1}^{L}\prod_{k=1}^{K}\frac{\theta_{lk}^{\varrho_{lk}}}{\Gamma(\varrho_{lk})}\left(B_{lk}\right)^{\varrho_{lk}-1}e^{-\theta_{lk}B_{lk}} \\
& \times \prod_{l=1}^{L}\frac{b_l^{a_l}}{\Gamma(a_l)}\left(E_l\right)^{a_l-1}e^{-b_lE_l} \times \prod_{l=1}^{L}\frac{\kappa_l^{\iota_l}}{\Gamma(\iota_l)}\left(A_l\right)^{\iota_l-1}e^{-\kappa_lA_l}
\end{aligned}
\tag{252}
$$

The complete set of parameters necessary for the model shown by $\Theta = \{z^{(l)}, z^{(g)}, \beta^{(l)}, \beta^{(g)}, \theta^{(g)}, \theta^{(l)}, \mu, B, \sigma, E,$

Fig. 6.2 provides a graphical representation of the model.

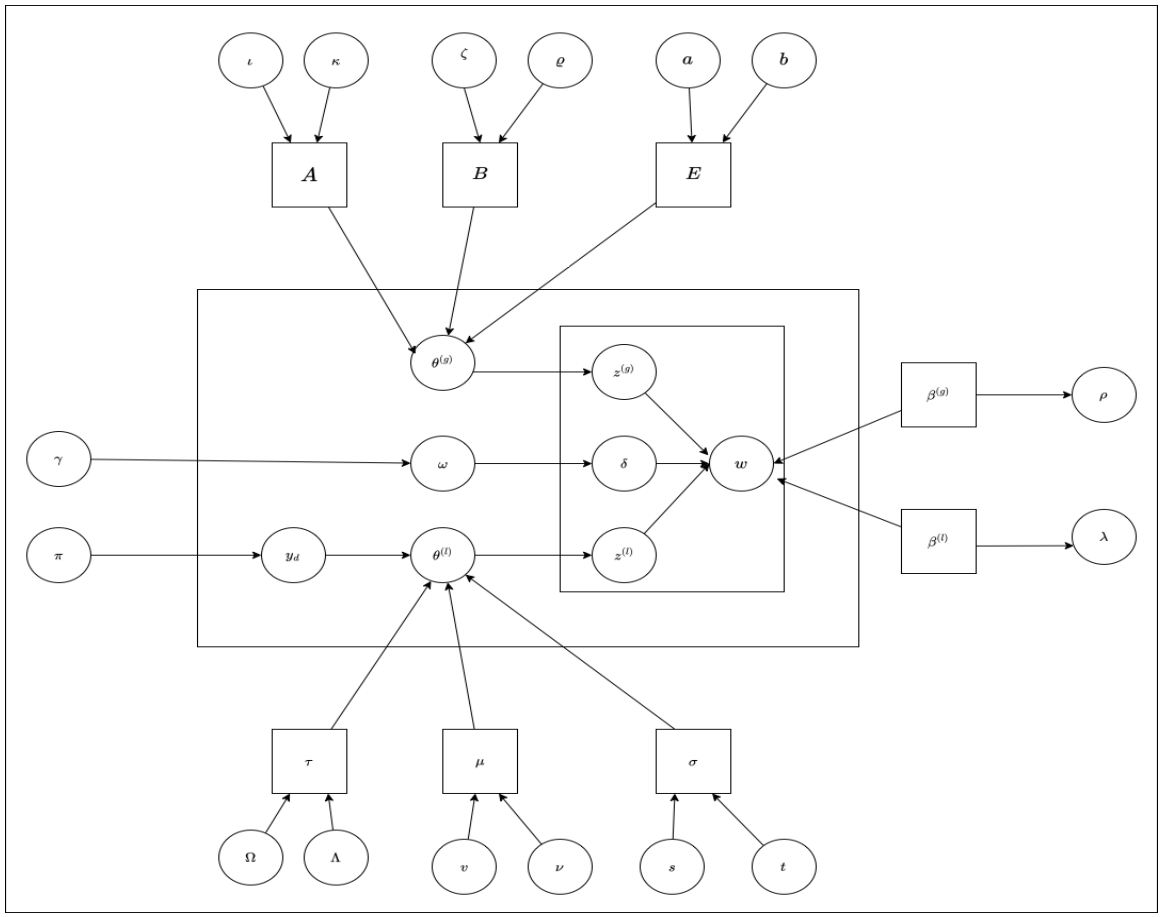


Figure 6.2: Graphical representation of MGBLBM

6.3.1 Variational solutions for MGBLBM

The variational solutions for Equation 252 are largely similar to those in the previous section, with some differences in the definitions of variables and the apparent change in $Q(\vec{\theta})$. Table 6.2 summarizes the relevant variables for the MGBLBM model, and the variational solutions are as follows:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D} \prod_{l=1}^{L} r_{dl}^{y_{dl}} \tag{253}$$

Table 6.2: Definitions of Parameters Used in the Equations MGBLBM model

D	Number of documents
L	Number of document groups
N_d	Number of words in document d
K	Number of topics
V	Vocabulary size
r_{dl}	Document-group assignment probability for document d and group l
y_{dl}	Indicator variable if document d is assigned to group l
$\phi_{dnk}^{(g)}$	Global topic assignment probability for word n in document d and topic k
$z_{dnk}^{(g)}$	Indicator variable if word n in document d is assigned to global topic k
$\phi_{dnk}^{(l)}$	Local topic assignment probability for word n in document d and topic k
$z_{dnk}^{(l)}$	Indicator variable if word n in document d is assigned to local topic k
σ_{lk}, τ_{lk}	Parameters for the local topic proportions distribution
$\vartheta_{lk}^*, \nu_{lk}^*$	Hyperparameters for the prior distribution of σ_{lk}
t_{lk}^*, s_{lk}^*	Hyperparameters for the prior distribution of τ_{lk}
$\Omega_{kv}^{(l)}$	Local word-topic probability for word v given topic k
λ_{kv}^*	Hyperparameters for the prior distribution of $\Omega_{kv}^{(l)}$
α_{lk}, β_{lk}	Parameters for the global topic proportions distribution
h_{lk}^*, g_{lk}^*	Hyperparameters for the prior distribution of α_{lk}
a_{lk}^*, b_{lk}^*	Hyperparameters for the prior distribution of β_{lk}
$\Omega_{kv}^{(g)}$	Global word-topic probability for word v given topic k
κ_{kv}^*	Hyperparameters for the prior distribution of $\Omega_{kv}^{(g)}$
$\theta_{dk}^{(l)}$	Local topic proportion for document d and topic k
$\iota_{dk}^*, \varrho_{dk}^*$	Hyperparameters for the local topic proportions
$\theta_{dk}^{(g)}$	Global topic proportion for document d and topic k
χ_{dk}^*, s_{dk}^*	Hyperparameters for the global topic proportions
$\rho_{dl}, \delta_{dnk}^{(l)}, \delta_{dnk}^{(g)}$	Intermediate variables for probability calculations
μ_{lk}	Intermediate variable for the prior distribution
B_{lk}	Parameter related to $\mathcal{R}^{(g)}$
E_l, A	Parameters for the Taylor series approximations
$\mathcal{R}^{(l)}, \mathcal{R}^{(g)}$	Taylor series approximations for the digamma and trigamma functions

$$Q(z^{(l)}) = \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K}\phi_{dnk}^{(g)z_{dnk}^{(l)}}, \quad Q(z^{(g)}) = \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K}\phi_{dnk}^{(g)z_{dnk}^{(g)}} \tag{254}$$

$$\begin{aligned} Q(\mu) &= \prod_{l=1}^{L}\prod_{k=1}^{K}\frac{{\nu_{lk}^*}^{\upsilon_{lk}^*}}{\Gamma(\upsilon_{lk}^*)}\mu_{lk}^{\upsilon_{lk}^*-1}e^{-\nu_{lk}^*\mu_{lk}}, \quad Q(\sigma_l) = \prod_{l=1}^{L}\frac{{t_l^*}^{s_l^*}}{\Gamma(s_l^*)}\sigma_l^{s_l^*-1}e^{-t_l^*\sigma_l}, \\ Q(\tau_l) &= \prod_{l=1}^{L}\frac{{\Lambda_l^*}^{\Omega_l^*}}{\Gamma(\Omega_l^*)}\tau_l^{\Omega_l^*-1}e^{-\Lambda_l^*\tau_l} \end{aligned} \tag{255}$$

$$\begin{aligned} Q(B) &= \prod_{l=1}^{L}\prod_{k=1}^{K}\frac{{\zeta_{lk}^*}^{\varrho_{lk}^*}}{\Gamma(\varrho_{lk}^*)}B_{lk}^{\varrho_{lk}^*-1}e^{-\zeta_{lk}^*B_{lk}}, \quad Q(E_l) = \prod_{l=1}^{L}\frac{{b^*}^{a_l^*}}{\Gamma(a_l^*)}E_l^{a_l^*-1}e^{-b^*E_l} \\ Q(A) &= \prod_{l=1}^{L}\frac{{\kappa_l^*}^{\iota_l^*}}{\Gamma(\iota_l^*)}A^{\iota_l^*-1}e^{-\kappa_l^*A} \end{aligned} \tag{256}$$

$$\begin{aligned} Q(\beta^{(l)}) &= \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma(\sum_{v=1}^{V}\lambda_{kv}^*)}{\prod_{v=1}^{V}\Gamma(\lambda_{kv}^*)}\beta_{kv}^{(l)\lambda_{kv}^*-1}, \\ Q(\beta^{(g)}) &= \prod_{k=1}^{K}\prod_{v=1}^{V}\frac{\Gamma(\sum_{v=1}^{V}\rho_{kv}^*)}{\prod_{v=1}^{V}\Gamma(\rho_{kv}^*)}\beta_{kv}^{(g)\rho_{kv}^*-1} \end{aligned} \tag{257}$$

$$\begin{aligned} Q(\theta^{(l)}) = &\prod_{d=1}^{D}\prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}f_{dk}^*)}{\Gamma(f_{dk}^*)}\frac{\Gamma(g_d^*+h_d^*)}{\Gamma(g_d^*)\Gamma(h_d^*)}\theta_{dk}^{f_{dk}^*-1} \\ &\times\Big[\sum_{k=1}^{K}\theta_{dk}^{(l)}\Big]^{g_d^*-\sum_{k=1}^{K}f_{dk}^*}\Big[1-\sum_{k=1}^{K}\theta_{dk}^{(l)}\Big]^{h_d^*-1} \end{aligned} \tag{258}$$

$$\begin{aligned} Q(\theta^{(g)}) = &\prod_{d=1}^{D}\prod_{k=1}^{K}\frac{\Gamma(\sum_{k=1}^{K}c_{dk}^*)}{\Gamma(c_{dk}^*)}\frac{\Gamma(R_d^*+Q_d^*)}{\Gamma(R_d^*)\Gamma(Q_d^*)}\theta_{dk}^{(g)c_{dk}^*-1} \\ &\times\Big[\sum_{k=1}^{K}\theta_{dk}^{(g)}\Big]^{R_d^*-\sum_{k=1}^{K}c_{dk}^*}\Big[1-\sum_{k=1}^{K}\theta_{dk}^{(g)}\Big]^{Q_d^*-1} \end{aligned} \tag{259}$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dnk}^{(l)} = \frac{\delta_{dnk}^{(l)}}{\sum_{k=1}^{K} \delta_{dnk}^{(l)}}, \pi_l = \frac{1}{D} \sum_{d=1}^{D} r_{dl}, \phi_{dnk}^{(g)} = \frac{\delta_{dnk}^{(g)}}{\sum_{k=1}^{K} \delta_{dnk}^{(g)}} \tag{260}$$

$$\begin{aligned} \rho_{dl} = \exp \Bigg\{ & \ln \pi_l + \mathcal{R}_l + \mathcal{S}_l + (\mu_{lk} - 1) \ln \theta_{dk}^{(l)} \\ & + \left(\sigma_l - \sum_{k=1}^{K} \mu_{lk} \right) \ln \left[\sum_{k=1}^{K} \theta_{dk}^{(l)} \right] \\ & + (\tau_l - 1) \ln \left[1 - \sum_{k=1}^{K} \theta_{dk}^{(l)} \right] \Bigg\} \end{aligned} \tag{261}$$

$$\delta_{dnk}^{(g)} = \exp(\ln \beta_{kv}^{(g)} + \ln \theta_{dk}^{(g)}) \tag{262}$$

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

$$\begin{aligned} \mathcal{R}^{(l)} = & \ln \frac{\Gamma(\sum_{k=1}^{K} \mu_{lk})}{\prod_{k=1}^{K} \Gamma(\mu_{lk})} + \sum_{k=1}^{K} \overline{\mu}_{lk} \Big[\Psi\Big(\sum_{k=1}^{K} \overline{\mu}_{lk} \Big) - \Psi(\overline{\mu}_{lk}) \Big] \big[\ln \mu_{lk} - \ln \overline{\mu}_{lk} \big] \\ & + \frac{1}{2} \sum_{k=1}^{K} \overline{\mu}_{lk}^2 \Big[\Psi'\big(\sum_{k=1}^{K} \overline{\mu}_{lk} \big) - \Psi'(\overline{\mu}_{lk}) \Big] - (\ln \mu_{lk} - \ln \overline{\mu}_{lk})^2 \\ & + \frac{1}{2} \sum_{a=1}^{K} \sum_{b=1, a \neq b}^{K} \overline{\mu}_{la} \overline{\mu}_{lb} \Big[\Psi'\Big(\sum_{k=1}^{K} \overline{\mu}_{lk} \Big) \big(\ln \mu_{la} - \ln \overline{\mu}_{la} \big) \big(\ln \mu_{lb} - \ln \overline{\mu}_{lb} \big) \Big] \end{aligned} \tag{263}$$

$$\begin{aligned} \mathcal{R}^{(g)} = & \ln \frac{\Gamma(\sum_{k=1}^{K} B_{lk})}{\prod_{k=1}^{K} \Gamma(B_{lk})} + \sum_{k=1}^{K} \overline{B}_{lk} \Big[\Psi\Big(\sum_{k=1}^{K} \overline{B}_{lk} \Big) - \Psi(\overline{B}_{lk}) \Big] \big[\ln B_{lk} - \ln \overline{B}_{lk} \big] \\ & + \frac{1}{2} \sum_{k=1}^{K} \overline{B}_{lk}^2 \Big[\Psi'\big(\sum_{k=1}^{K} \overline{B}_{lk} \big) - \Psi'(\overline{B}_{lk}) \Big] - (\ln B_{lk} - \ln \overline{B}_{lk})^2 \\ & + \frac{1}{2} \sum_{a=1}^{K} \sum_{b=1, a \neq b}^{K} \overline{B}_{la} \overline{B}_{lb} \Big[\Psi'\Big(\sum_{k=1}^{K} \overline{B}_{lk} \Big) \big(\ln B_{la} - \ln \overline{B}_{la} \big) \big(\ln B_{lb} - \ln \overline{B}_{lb} \big) \Big] \end{aligned} \tag{264}$$

$$
\begin{aligned}
\mathcal{S}^{(l)} =& \ln \frac{\Gamma(\overline{\sigma}+\overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\sigma})\big](\ln\sigma - \ln\overline{\sigma}) \\
&+ \overline{\tau}\big[\Psi(\overline{\sigma}+\overline{\tau}) - \Psi(\overline{\tau})\big](\ln\tau - \ln\overline{\tau}) \\
&+ 0.5\overline{\sigma}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\sigma})\big](\ln\sigma - \ln\overline{\sigma})^2 \\
&+ 0.5\overline{\tau}^2\big[\Psi'(\overline{\sigma}+\overline{\tau}) - \Psi'(\overline{\tau})\big](\ln\tau - \ln\overline{\tau})^2 \\
&+ \overline{\sigma}\,\overline{\tau}\,\Psi'(\overline{\sigma}+\overline{\tau})(\ln\sigma - \ln\overline{\sigma})(\big\langle \ln\tau\big\rangle - \ln\overline{\tau})
\end{aligned}
\tag{265}
$$

$$
\begin{aligned}
\mathcal{S}^{(\})} =& \ln \frac{\Gamma(\overline{E}+\overline{A})}{\Gamma(\overline{E})\Gamma(\overline{A})} + \overline{E}\big[\Psi(\overline{E}+\overline{A}) - \Psi(\overline{E})\big](\ln E - \ln\overline{E}) \\
&+ \overline{A}\big[\Psi(\overline{E}+\overline{A}) - \Psi(\overline{A})\big](\ln A - \ln\overline{A}) \\
&+ 0.5\overline{E}^2\big[\Psi'(\overline{E}+\overline{A}) - \Psi'(\overline{E})\big](\ln E - \ln\overline{E})^2 \\
&+ 0.5\overline{A}^2\big[\Psi'(\overline{E}+\overline{A}) - \Psi'(\overline{A})\big](\ln A - \ln\overline{A})^2 \\
&+ \overline{E}\,\overline{A}\,\Psi'(\overline{E}+\overline{A})(\ln E - \ln\overline{E})(\big\langle \ln A\big\rangle - \ln\overline{A})
\end{aligned}
\tag{266}
$$

$$
\begin{aligned}
\upsilon^*_{lk} =& \upsilon_{lk} + \sum_{d=1}^{D} y_{dl}\overline{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K}\overline{\mu}_{lk}\Big) - \Psi(\overline{\mu}_{lk}) \\
&+ \Psi\Big(\sum_{k=1}^{K}\Big)\sum_{a\neq k}^{K}\big(\ln\mu_{la} - \ln\overline{\mu}_{la}\big)\overline{\mu}_{la}\Big]
\end{aligned}
\tag{267}
$$

$$
\nu^*_{lk} = \nu_{lk} - \sum_{d=1}^{D} y_{dl}\Big[\ln\theta_{dk} - \ln\sum_{k=1}^{K}\theta_{dk}\Big] \tag{268}
$$

$$
\begin{aligned}
s^*_l =& s_l + \sum_{d=1}^{D} y_{dl}\Bigg[\Psi\big(\overline{\sigma}_l + \overline{\tau}_l\big) - \Psi\big(\overline{\sigma}_l\big) \\
&+ \overline{\tau}_l\Psi'\big(\overline{\sigma}_l + \overline{\tau}_l\big)\big(\ln\tau_l - \ln\overline{\tau}_l\big)\Bigg]\overline{\sigma}_l
\end{aligned}
\tag{269}
$$

$$t_l^* = t_l - \sum_{d=1}^{D} y_{dl} \Big\langle \ln \Big[\sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{270}$$

$$\begin{aligned} \Omega_l^* =& \Omega_{lk} + \sum_{d=1}^{D} y_{dl} \Big[\Psi\big(\overline{\tau}_l + \overline{\sigma}_l\big) - \Psi\big(\overline{\tau}_l\big) \\ &+ \overline{\sigma}_l \Psi'\big(\overline{\tau}_l + \overline{\sigma}_l\big)\big(\ln \sigma_l - \ln \overline{\sigma}_l\big) \Big] \overline{\tau}_l \end{aligned} \tag{271}$$

$$\Lambda_l^* = \Lambda_l - \sum_{d=1}^{D} y_{dl} \Big\langle \ln \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big] \Big\rangle \tag{272}$$

$$\begin{aligned} f_{dk}^* = f_{dk} + \sum_{n=1}^{N_d} z_{dnk} + \sum_{l=1}^{L} y_{dl} \mu_{lk}, \quad g_d^* = g_d + \sum_{n=1}^{N_d} \sum_{k=1}^{K} z_{dnk} + \sum_{l=1}^{L} y_{dl}, \\ h_d^* = h_d + \sum_{l=1}^{L} y_{dl} \tau_l \end{aligned} \tag{273}$$

$$\begin{aligned} \varrho_{lk}^* =& \varrho_{lk} + \overline{B}_{lk} \Big[\Psi\Big(\sum_{k=1}^{K} \overline{B}_{lk} \Big) - \Psi(\overline{B}_{lk}) \\ &+ \Psi\Big(\sum_{k=1}^{K} \Big) \sum_{a \neq k}^{K} \big(\ln B_{la} - \ln \overline{B}_{la}\big) \overline{B}_{la} \Big] \end{aligned} \tag{274}$$

$$\zeta_{lk}^* = \zeta_{lk} - \Big[\ln \theta_{dk}^{(g)} - \ln \sum_{k=1}^{K} \theta_{dk}^{(g)} \Big] \tag{275}$$

$$a_l^* = a_l + \Psi\big(\overline{E}_l + \overline{A}_l\big) - \Psi\big(\overline{E}_l\big) + \overline{A}_l \Psi'\big(\overline{E}_l + \overline{A}_l\big)\big(\ln A - \ln \overline{A}_l\big) \overline{E}_l \tag{276}$$

$$b^* = b - \sum_{d=1}^{D} \Big\langle \ln \Big[\sum_{k=1}^{K} \theta_{dk}^{(g)} \Big] \Big\rangle \tag{277}$$

$$\iota_l^* = \iota_{lk} + \sum_{d=1}^{D} \Big[\Psi\big(\overline{A}_l + \overline{E}_l\big) - \Psi\big(\overline{A}_l\big) + \overline{E}_l \Psi'\big(\overline{A}_l + \overline{E}_l\big)\big(\ln E_l - \ln \overline{E}_l\big) \Big] \overline{A}_l \tag{278}$$

$$\kappa_l^* = \kappa_l - \sum_{d=1}^{D} \left\langle \ln \left[1 - \sum_{k=1}^{K} \theta_{dk}^{(g)} \right] \right\rangle \tag{279}$$

$$c_{dk}^* = c_{dk} + \sum_{n=1}^{N_d} z_{dnk} + \sum_{l=1}^{L} B_{lk} \tag{280}$$

$$R_d^* = R_d + \sum_{n=1}^{N_d} \sum_{k=1}^{K} z_{dnk} + \sum_{l=1}^{L} E_l \tag{281}$$

$$Q_d^* = Q_d + \sum_{l=1}^{L} A \tag{282}$$

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

$$\ln \theta_{dk}^{(l)} = \Psi(f_{dk}) - \Psi\Big(\sum_{k=1}^{K} f_{dk} \Big) + \Psi(g_d) - \Psi(g_d + h_d) \tag{283}$$

$$\sum_{k=1}^{k} \theta_{dk}^{(l)} = \sum_{k=1}^{k} \big(\Psi(g_d) - \Psi(g_d + h_d) \big) \tag{284}$$

$$1 - \sum_{k=1}^{k} \theta_{dk}^{(l)} = \sum_{k=1}^{k} \big(\Psi(h_d) - \Psi(g_d + h_d) \big) \tag{285}$$

$$\overline{\sigma}_{lk} = \frac{\upsilon_{lk}^*}{\nu_{lk}^*}, \ln \sigma_{lk} = \Psi\big(\upsilon_{lk}^*\big) - \ln \nu_{lk}^* \tag{286}$$

$$\big(\ln \sigma_{lk} - \ln \overline{\sigma}_{lk} \big)^2 = \big[\Psi\big(\upsilon_{lk}^*\big) - \ln \upsilon_{lk}^* \big]^2 + \Psi'\big(\upsilon_{lk}^*\big) \tag{287}$$

$$\overline{\sigma}_l = \frac{s_l^*}{t_l^*}, \ln \sigma_l = \Psi\big(s_l^*\big) - \ln t_l^* \tag{288}$$

$$\big(\ln \sigma_l - \ln \overline{\sigma}_l \big)^2 = \big[\Psi\big(s_l^*\big) - \ln s_l^* \big]^2 + \Psi'\big(s_l^*\big) \tag{289}$$

$$\overline{\tau}_{lk} = \frac{\Omega_l^*}{\Lambda_l^*}, \ln \tau_l = \Psi\big(\Omega_l^*\big) - \ln \Lambda_l^* \tag{290}$$

$$\big(\ln \tau_l - \ln \overline{\tau}_l \big)^2 = \big[\Psi\big(\Omega_l^*\big) - \ln \Omega_l^* \big]^2 + \Psi'\big(\Omega_l^*\big) \tag{291}$$

$$z_{dnk}^{(l)} = \phi_{dnk}^{(l)}, y_{dl} = r_{dl}, \ln \beta_{kv}^{(l)} = \Psi(\lambda_{kv}) - \Psi(\sum_{f=1}^{V} \lambda_{kf}) \tag{292}$$

$$\ln \theta_{dk}^{(g)} = \Psi(c_{dk}) - \Psi\Big(\sum_{k=1}^{K} c_{dk}\Big) + \Psi(R_d) - \Psi(R_d + Q_d) \tag{293}$$

$$\sum_{k=1}^{k} \theta_{dk}^{(g)} = \sum_{k=1}^{k} \big(\Psi(R_d) - \Psi(R_d + Q_d)\big) \tag{294}$$

$$1 - \sum_{k=1}^{k} \theta_{dk}^{(g)} = \sum_{k=1}^{k} \big(\Psi(Q_d) - \Psi(R_d + Q_d)\big) \tag{295}$$

$$\overline{E}_{lk} = \frac{\varrho_{lk}^*}{\zeta_{lk}^*}, \ln E_{lk} = \Psi\big(\varrho_{lk}^*\big) - \ln \zeta_{lk}^* \tag{296}$$

$$\big(\ln E_{lk} - \ln \overline{E}_{lk}\big)^2 = \big[\Psi\big(\varrho_{lk}^*\big) - \ln \varrho_{lk}^*\big]^2 + \Psi'\big(\varrho_{lk}^*\big) \tag{297}$$

$$\overline{E}_{l} = \frac{a_{l}^*}{b^*}, \ln E_{l} = \Psi\big(a_{l}^*\big) - \ln b^* \tag{298}$$

$$\big(\ln E_{l} - \ln \overline{E}_{l}\big)^2 = \big[\Psi\big(a_{l}^*\big) - \ln a_{l}^*\big]^2 + \Psi'\big(a_{l}^*\big) \tag{299}$$

$$\overline{A}_{lk} = \frac{\iota_{l}^*}{\kappa_{l}^*}, \ln A = \Psi\big(\iota_{l}^*\big) - \ln \kappa_{l}^* \tag{300}$$

$$\big(\ln A - \ln \overline{A}_{l}\big)^2 = \big[\Psi\big(\iota_{l}^*\big) - \ln \iota_{l}^*\big]^2 + \Psi'\big(\iota_{l}^*\big) \tag{301}$$

$$z_{dnk}^{(g)} = \phi_{dnk}^{(g)}, \ln \beta_{kv}^{(g)} = \Psi(\rho_{kv}) - \Psi(\sum_{f=1}^{V} \rho_{kf}) \tag{302}$$

We follow the same algorithm for the LBLMA, calculating equations 253 - 259 repeatedly until convergence.

The optimization of the model parameters was achieved by maximizing the lower bound, as demonstrated in the following:

$$\beta^{(l)} = \sum_{d=1}^{D} \sum_{n=2}^{N_d} \sum_{v=1}^{V} f * g * h\theta_{dnk}^{(l)} w_{dnv} \tag{303}$$

$$\beta^{(g)} = \sum_{d=1}^{D} \sum_{n=2}^{N_d} \sum_{v=1}^{V} c * R * Q\theta_{dnk}^{(g)} w_{dnv} \tag{304}$$

Beta-Liouville Parameter

The goal of this subsection is to find the model's parameter estimates based on variational inferences.

$$\begin{aligned}\mathcal{L}[\theta] = &\sum_{m=1}^{M}(\log(\Gamma(\sum_{l=1}^{D}c_l)) + \log(\Gamma(R+Q)) - \log\Gamma(R) - \log\Gamma(Q) \\ &- \sum_{i=1}^{D}\log\Gamma(c) + \sum_{i=1}^{D}c(\Psi(\gamma_{mi}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) \\ &+ R(\Psi(C_{m\gamma}) - \Psi(C_{m\gamma}Q_{m\gamma})) + Q(\Psi(Q_{m\gamma}) - \Psi(C_{m\gamma}+Q_{m\gamma})))\end{aligned} \tag{305}$$

The derivative of the above equation with respect to the BL parameter is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}[\theta]}{\partial c} &= M(\Psi(\sum_{l=1}^{D}) - \Psi(c)) + \sum_{m=1}^{M}(\Psi^{'}(\gamma_{ml}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) \\ \frac{\partial\mathcal{L}[\theta]}{\partial R} &= M[\Psi(R+Q) - \Psi(R)] + \sum_{m=1}^{M}(\Psi(c_{m\gamma}) - \Psi(c_{m\gamma}+Q_{m\gamma})) \\ \frac{\partial\mathcal{L}[\theta]}{\partial Q} &= M[\Psi(R+Q) - \Psi(Q)] + \sum_{m=1}^{M}(\Psi(Q_{m\gamma}) - \Psi(R_{m\gamma}+Q_{m\gamma}))\end{aligned} \tag{306}$$

It is clear from the preceding equations that the derivative in Eq. 305 with respect to each of the BL parameters depends not only on their own values but also on each other. As a result, we employ the Newton-Raphson method to solve the optimization problem. To use the Newton-Raphson method, we must first compute the Hessian matrix with respect to the parameter space, as shown below:

$$\begin{aligned}\frac{\partial^2\mathcal{L}[\theta]}{\partial cR_j} &= M(-\delta(i,j)\Psi^{'}(c) + \Psi^{'}(\sum_{l=1}^{D}c)) \\ \frac{\partial^2\mathcal{L}[\theta]}{\partial R^2} &= M(\Psi^{'}(R+Q) - \Psi^{'}(R)) \\ \frac{\partial^2\mathcal{L}[\theta]}{\partial R\partial Q} &= M\Psi^{'}(R+Q) \\ \frac{\partial^2\mathcal{L}[\theta]}{\partial Q^2} &= M(\Psi^{'}(R+Q) - \Psi^{'}(Q))\end{aligned} \tag{307}$$

The Hessian matrix shown above is very similar to the Hessian matrix of the Dirichlet parameters in the MPCA model and generalized Dirichlet parameters in GDMPCA. In fact, the above matrix

can be divided into two completely separate matrices by the parameters c, R and Q. Each of the two parts' parameter derivation will be identical to the Newton-Raphson model provided by MPCA and GDMPCA.

6.4 Experimental Results

In this section, we assess the effectiveness of our proposed algorithms through two rigorous applications: topic modeling for medical texts and sentiment analysis. We evaluate each model by examining its success rate for each dataset and its perplexity [37, 10, 4, 101], a standard metric in language modeling, defined as follows:

$$prep(\mathcal{D}_{test}) = \exp\left(\frac{-\ln p(\mathcal{D}_{test})}{\sum_d |w_d|}\right) \tag{308}$$

where d is the length of the page and $|w_d|$ is its width. A lower perplexity score indicates better generalization performance. In addition to evaluating the perplexity metric, we also consider time complexity and likelihood to assess our models. Our research aims to evaluate a variety of models including MGMLMB, MGGDCTM, MGCTM, CTM, LDA, and NMF, to identify the local topic. The datasets selected are critical for our study as they encompass a wide range of analytical scenarios. From topic modeling in medical texts to broader applications, these datasets allow for an in-depth evaluation of the models' flexibility and precision. By analyzing datasets with unique features, we can showcase the capabilities of our proposed models in different settings, emphasizing their utility as a comprehensive tool in text analysis.

6.4.1 Topic Modeling for Medical Texts

The primary objective of text classification is to methodically categorize various documents into predefined subject categories, as highlighted by earlier research [102]. This domain has been thoroughly examined, yielding a wide array of solutions [103, 46]. Topic modeling, a prevalent technique in natural language processing, stands out as a particularly effective method. It offers versatility in analyzing diverse texts, from news articles and tweets to creating visual representations of related topics and documents. Additionally, topic modeling addresses the challenges of high

dimensionality and data sparsity, commonly encountered in sectors like health and medical text mining, despite the large volumes of data available [107]. Initially developed for text analysis where documents are analyzed based on the frequency of phrase usage, topic modeling involves a suite of statistical learning techniques that identify hidden or 'latent' topics in extensive text data without direct supervision. In this context, a 'topic' is defined as a cluster of keywords that follows a probability distribution, and a 'document' consists of a mix of such topics, adhering to a similar distribution pattern.

According to [13], topic modeling primarily produces a set of keywords per topic, which proves to be highly effective especially in health and medical research by improving the extraction and comprehension of essential data insights. However, the abundance of data still necessitates further advancements in topic modeling methods [107]. In the field of biological natural language processing (BioNLP) [150], topic models are particularly advantageous. They enhance the processing and understanding of complex, domain-specific biological texts, thereby boosting information retrieval and aiding in more precise scientific investigations. Additionally, topic modeling in biology helps reveal underlying thematic structures or latent topics, which may lead to the discovery of new correlations and insights. This function is also crucial for systematically organizing and categorizing vast datasets, making biological data more manageable and easier to analyze for both researchers and professionals in the industry.

To evaluate our models, we chose the medical transcription dataset [130], mental health dataset [131], and Genia dataset [132].

Mental Health Tweet

The Mental Health Corpus [131] contains textual data related to mental health conditions like anxiety and depression, structured into discussions and categorized as toxic or non-toxic. This dataset is valuable for sentiment analysis, detecting harmful language, and studying language patterns in mental health discourse. It is beneficial for researchers, practitioners, and those interested in mental health discourse. Our study focused on detecting mental health conditions in the text using topic modeling. This statistical method helps identify prevalent themes in the data, enhancing our understanding of mental health discussions.

Table 6.3 displays the initial five local topics identified using the MGDCTM.

Table 6.3: Common topics identified with MGDCTM model in the Mental Health Tweet dataset, each defined by a set of keywords

Topic No	Topics
Topic1	'alcohol', 'touched', 'impulse', 'happened', 'learn', 'honest', 'follow', 'obvious', 'situation', 'dizziness'
Topic2	'friend', 'injury', 'anxiety', 'follow', 'worried', 'happened', 'lowcost', 'groups', 'trapped', 'pharmacy'
Topic3	'suicidal', 'receive', 'obvious', 'pharmacy', 'events', 'member', 'injury', 'network', 'pattern', 'counselor'
Topic4	'network', 'dumbbell', 'pharmacy', 'matter', 'unlike', 'evidence', 'response', 'critical', 'services', 'swapping'
Topic5	'health', 'doctor's', 'employee', 'selfhelp', 'working', 'attach', 'online', 'influence', 'wellbeing', 'health'

Fig. 6.3 illustrates the perplexity of various topic modeling algorithms applied to the Mental Health Tweet dataset across different numbers of topics. LDA, NMF, MGCTM and CTM exhibit increasing perplexity with more topics, suggesting difficulty in handling larger topic numbers. In contrast, MGDCTM and MBLCTM maintain lower perplexity, indicating better performance. These results suggest that MGDCTM, MBLCTM, and MGCTM might be more effective for analyzing complex data patterns in mental health discussions on social media.

Table 6.4 shows the time complexity in minutes for various topic modeling algorithms on the Mental Health Tweet dataset across topic counts. Generally, as the number of topics increases, so does the computational time for each model, highlighting the increasing demands of handling more complex topic structures. This provides insight into each model's efficiency at scaling with larger datasets.

Table 6.5 presents the likelihood values for different topic modeling algorithms applied to the Mental Health Tweet dataset, with topic counts K=5, 10, 15, 20 and 25. These likelihood values measure how well each model fits the data; a value closer to zero indicates a better fit. As the number of topics increases, the likelihood values become more negative, suggesting that model fit generally decreases with more complex models across most algorithms, highlighting the challenges of modeling larger topic spaces effectively.

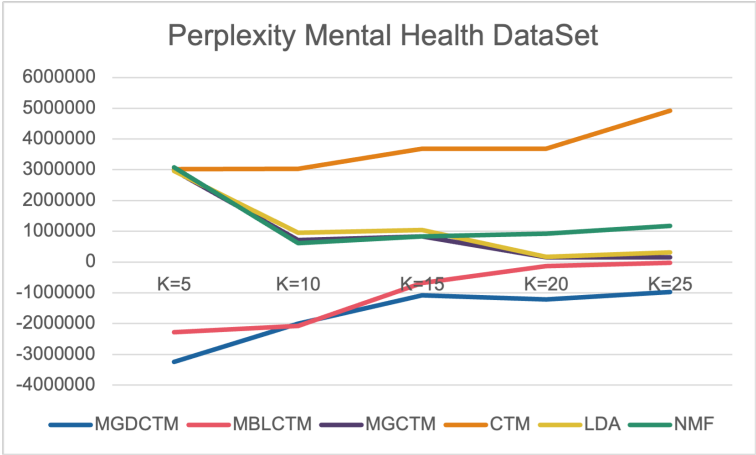


Figure 6.3: Perplexity for Mental Health Tweet dataset

Table 6.4: Time complexity comparison for for different model at varying topic levels (K) o Mental Health Tweet dataset (min)

K	5	10	15	20	25
NMF	2.05	2.37	2.87	3.075	3.239
LDA	1.92	2.496	2.688	2.88	3.0336
CTM	1.4712	1.6812	2.05968	2.2068	2.324496
MGCTM	1.0512	1.314	1.7082	1.7739	2.12868
MGDCTM	1.7226	1.98	2.6136	2.6928	2.772
MBLCTM	1.3008	1.084	1.4092	1.874236	2.2490832

Genia Dataset

Texts in the field of biomedicine are invaluable resources for advancing medical knowledge. Traditionally, extracting information from these texts required manual effort by domain experts, but automation can dramatically accelerate progress in medical research. Biomedical texts, for example, can reveal how drugs affect individuals and help diagnose health conditions. Thus, automated event extraction from biomedical texts is extremely useful. This involves identifying the original text, annotating trigger words, determining their exact locations within the text, and classifying the types of events they signify. Automating this process enhances both the efficiency and accuracy of information extraction from biomedical literature, facilitating quicker advancements in the medical field [132].

The MGDCTM identified the initial five local topics, which are displayed in Table 6.6.

Table 6.5: Likelihood comparison for different topic models approaches on Mental Health Tweet dataset

K	5	10	15	20	25
NMF	-122328887	-17040338.76	-32934820.18	-42730466.24	-81393075.8
LDA	-122341234.32	-17127772.76	-33032133.18	-44042844.24	-85638863.8
CTM	-82689087.2	-21424478.76	-30271575.68	-46517304.52	-49781525.44
MGCTM	-70343409.2	-21103022.76	-28137363.68	-42206045.52	-49240386.44
MGDCTM	-1013906442	-1165409704	-2074429273	-1538340809	-1619919488
MBLCTM	-2527735331	-1209442742	-2297941210	-2987323573	-3943267117

Table 6.6: Common topics identified with MGDCTM model in the Genia dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'syndromes', 'increases', 'renilla', 'effector', 'repressed', 'retinoid', 'induced', 'supported', 'ferritin', 'control'
Topic 2	'plasmids', 'inducible', 'objective', 'inhibited', 'resides', 'presented', 'complex', 'activity', 'domains', 'showed'
Topic 3	'plasmids', 'visible', 'resides', 'inducible', 'challenge', 'events', 'requires', 'latently', 'mediate', 'replaced'
Topic 4	'regulator', 'removed', 'distinct', 'methyl', 'aggarwal', 'thtype', 'little', 'proximal', 'expresses', 'spread'
Topic 5	'teflon', 'methyl', 'increased', 'finger', 'activates', 'perforin', 'depends', 'lectin', 'enhancer', 'species'

Fig 6.4 displays the perplexity scores for various topic modeling algorithms applied to the Genia dataset across different numbers of topics (K=5, 10, 15, 20, 25). The MGDCTM shows relatively stable and low perplexity across the board. The MBLCTM exhibits a significant decrease in perplexity as topic numbers increase. In contrast, the MGCTM and CTM demonstrate fluctuating scores, with a general increase as more topics are added. Both LDA and NMF show increases in perplexity, particularly at higher topic counts, indicating challenges in effectively modeling the complexity of the dataset.

Table 6.7 presents a likelihood comparison for different topic modeling approaches applied to the Genia dataset, across varying numbers of topics. The results show that as the number of topics increases, the likelihood values generally become more negative for all models, indicating a decrease in model performance with higher topic complexities.

Table 6.8 compares the computation times of several topic modeling techniques (NMF, LDA,

CTM, MGCTM, MGDCTM, and MBLCTM) as the number of topics increases. MBLCTM consistently shows the lowest computation times, indicating high efficiency, while NMF and LDA require more time, especially for larger topic counts. This information is key for selecting efficient models for processing large datasets like the Genia dataset.

Table 6.7: Likelihood comparison for different topic models approaches on Genia dataset

K	5	10	15	20	25
NMF	-122346665	-18360985.76	-33108452.18	-44166141.24	-86873386.8
LDA	-122341234	-17127772.76	-33032133.18	-44042844.24	-85638863.8
CTM	-85590799.54	-1734949.991	-4810783.977	-34271553.82	-5221871.972
MGCTM	-88806478.54	-1776129.571	-4440323.927	-35522591.42	-5328388.712
MGDCTM	-1091654960	-1080846495	-2053608340	-2669690842	-3523991912
MBLCTM	-1637211477	-1127556113	-1488374070	-1240311725	-1375618458

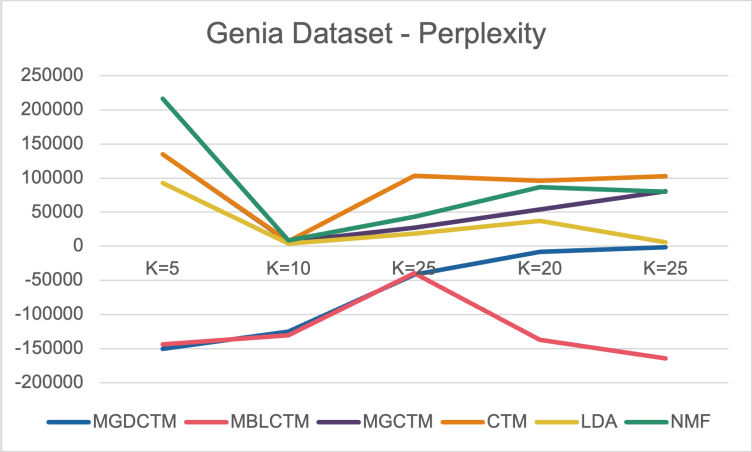


Figure 6.4: Perplexity for Genia dataset

Medical Transcription Dataset

Due to the stringent privacy regulations enforced by the Health Insurance Portability and Accountability Act (HIPAA), accessing comprehensive medical data can be particularly challenging. This often limits research and educational opportunities within the medical field. In response to this limitation, the MTSamples dataset emerges as a significant resource. It offers a diverse collection of medical transcription samples that encompass a wide array of medical specialties and employment contexts.

Table 6.8: Time complexity comparison for for different model at varying topic levels (K) on the Genia dataset. (min)

K	5	10	15	20	25
NMF	14.87	15.0187	16.9518	18.7362	18.8849
LDA	13.67	17.771	19.138	20.505	21.5986
CTM	10.74	13.962	15.036	16.11	16.9692
MGCTM	10.26	13.338	14.364	15.39	16.2108
MGDCTM	6.624	8.28	10.764	14.31612	17.179344
MBLCTM	6.705	7.45	9.685	10.0575	10.3555

The MTSamples dataset is meticulously curated to provide researchers, educators, and medical professionals with access to a substantial library of transcribed medical reports. These reports are categorized by specialty—ranging from cardiology to dermatology—and type of employment, enhancing their utility for specific research and training needs. Each report within the dataset has been transcribed by skilled transcriptionists and is provided exclusively for reference and educational purposes. This availability allows users to explore and utilize real-world medical data in a manner that adheres to privacy standards set forth by HIPAA, thereby facilitating a deeper understanding of medical documentation practices and terminology without compromising patient confidentiality [130]. Table 6.9 shows the first five local topics identified with the MGDCTM. Figure 6.5 presents the perplexity scores for various topic modeling algorithms applied to the Medical Transcription dataset across topic numbers ranging from K=5 to K=25. Perplexity, a measure of model prediction quality where lower scores indicate better model performance, varies across the models: MGDCTM shows consistently low perplexity, indicating strong model fit; MBLCTM starts low but increases at higher topic numbers, suggesting possible inefficiencies; MGCTM and CTM (Green Line) exhibit decreases in perplexity with increases in topics, potentially indicating improved fit; LDA remains relatively stable; and NMF shows moderate but increasing perplexity at higher topic counts.

Moreover, Table 6.10 provides a likelihood comparison of various topic modeling algorithms on the Medical Transcription dataset for different topic counts (K=5, 10, 15, 20, 25). These negative likelihood values show how well each model fits the data, with less negative values indicating a better fit. The models compared include NMF, LDA, CTM, MGCTM, MGDCTM, and MBLCTM.

Generally, as the number of topics increases, the likelihood values become more negative, suggesting reduced model performance with greater topic complexities. This table is useful for evaluating which models are best suited for analyzing complex datasets.

Table 6.11 shows the time complexity in minutes for various topic modeling algorithms on the Medical Transcription dataset across topic counts (K=5, 10, 15, 20, 25). The algorithms tested include NMF, LDA, CTM, MGCTM, MGDCTM, and MBLCTM. As the number of topics increases, the computational time generally rises for all models, highlighting the increasing complexity. Notably, MBLCTM exhibits the lowest time complexity across all topic sizes, indicating its efficiency. This table aids in evaluating each model's time efficiency, crucial for selecting the most suitable algorithm based on time constraints and computational resources.

Table 6.9: Common topics identified with MGDCTM model in theMedical Transcript dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'degreesc', 'ectopic', 'alphapal', 'abolished', 'memory', 'capable', 'complete', 'referred', 'scurfy', 'apobecf'
Topic 2	'hyperige', 'necrosis', 'membranes', 'isolated', 'abolished', 'mucida', 'remaining', 'product', 'nuclease', 'results'
Topic 3	'tonsil', 'identify', 'adenylate', 'present', 'fermentas', 'effect', 'nuclear', 'possibly', 'subject', 'provide'
Topic 4	'driven', 'control', 'developed', 'present', 'compared', 'stained', 'analyzed', 'retinoic', 'antimouse', 'peptide'
Topic 5	'rested', 'compared', 'antimouse', 'apospmut', 'peptide', 'ligated', 'region', 'sample', 'receptor', 'abolish'

Table 6.10: Likelihood comparison for different topic models approaches on Medical Transcription dataset

K	5	10	15	20	25
NMF	-12500109.68	-840911.041	-2600242.137	-9887223.478	-12613606.94
LDA	-12376650.91	-742599.131	-2475330.91	-8663655.84	-8044822.44
CTM	-1061612719	-202691745.7	-212980442.9	-394036705.2	-461917038.4
MGCTM	-528617780.5	-100437378.3	-105723556.1	-195588578.8	-227305645.6
MGDCTM	-1094492097	-1216102330	-2274111358	-1763348379	-1739026332
MBLCTM	-1123607107	-1291502422	-387450726.5	-516600968.6	-774901453

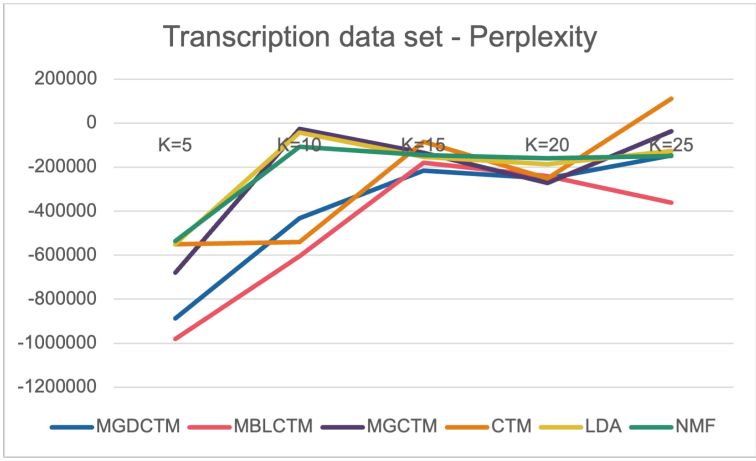


Figure 6.5: Perplexity for Medical Transcription dataset

Table 6.11: Time complexity comparison for for MPCA, GDMPCA, and BLMPCA at varying topic levels (K) on the Medical Transcription dataset (min)

K	5	10	15	20	25
NMF	4.9	5.02	5.23	7.028	7.53
LDA	4.67	6.071	6.538	7.005	7.3786
CTM	4.2672	4.4772	5.82036	6.04422	6.223308
MGCTM	3.5472	4.61136	4.96608	5.3208	5.604576
MGDCTM	2.748	3.435	4.4655	4.63725	5.5647
MBLCTM	2.2272	2.56	2.816	3.584	4.096

6.4.2 Topic Modeling

Topic modeling is extensively utilized for the purpose of clustering and managing vast collections of text data, which plays a significant role in the classification of textual content [104].

Considering the substantial number of documents available in data collections, it is impractical to analyze each document individually. To navigate this, one efficient approach involves identifying the key terms that define the corpus. This typically includes the most frequently occurring words. Alternatively, documents can be broken down into their constituent words and phrases. These elements are then grouped based on their similarities to form clusters. These clusters of words and phrases enable a deeper and more structured understanding of the underlying themes within the corpus. Essentially, the corpus is conceptualized as a collection of representative words, each selected from different clusters. This method stands in contrast to rule-based text mining techniques,

which depend on regular expressions and dictionary-based keyword searches. Instead, topic modeling endeavors to discover the pivotal words or themes in a text corpus autonomously, without prior assumptions [13].

To evaluate the performance and reliability of our proposed models, we utilized a comprehensive set of 2246 documents obtained from the Associated Press [13], providing a robust basis for our analysis.

Table 6.12 presents the first 5 local topics identified with the MGDCTM.

Table 6.13 compares perplexity scores for various topic modeling algorithms on the Associated Press dataset with topic counts from K=5 to K=25. Lower perplexity indicates better model performance. MGDCTM shows the lowest perplexity, suggesting the best fit, while NMF and LDA have the highest perplexity, indicating poorer performance. CTM and MGCTM display moderate perplexity, with MGCTM increasing significantly at higher topic counts. MBLCTM has lower perplexity than NMF and LDA but higher than MGDCTM, making it a moderately effective model. This table helps in selecting the most effective model for complex data analysis.

Furthermore, Fig. 6.6 displays the perplexity scores for various topic modeling algorithms on the Associated Press dataset across topic counts (K=5, 10, 15, 20, 25). Lower perplexity indicates better model performance. MGDCTM shows the lowest perplexity, indicating strong model fit, while MBLCTM also performs well but with slightly higher scores. MGCTM and CTM have moderate perplexity, increasing at higher topic counts. LDA and NMF exhibit higher perplexity, suggesting less effective performance. This graph helps identify which models are more effective for analyzing the dataset's complexity.

Moreover, Table 6.14 shows the time complexity, in minutes, for various topic modeling algorithms applied to the Associated Press dataset across different topic counts (K=5, 10, 15, 20, 25). The models include NMF, LDA, CTM, MGCTM, MGDCTM, and MBLCTM. As the number of topics increases, the time complexity rises for all models, with MGDCTM and MBLCTM showing the highest time complexity, especially at higher topic counts. NMF and LDA have relatively lower time complexities, making them more time-efficient compared to the others. This table highlights the computational demands of each model, aiding in the selection of efficient algorithms for large datasets.

Table 6.12: Common topics identified with MGDCTM model in the Associated Press dataset, each defined by a set of keywords

Topic No	Topics
Topic 1	'telephone', 'dealings', 'crumbling', 'killings', 'brightly', 'jersey', 'damico', 'belgium', 'martin', 'highway'
Topic 2	'fatigue', 'financial', 'jersey', 'guilders', 'enters', 'showing', 'session', 'takeover', 'security', 'candles'
Topic 3	'grocery', 'jersey', 'sentence', 'emotion', 'takeover', 'academy', 'texas', 'fatigue', 'telephone', 'nations'
Topic 4	'damico', 'crumbling', 'dealings', 'fruits', 'financial', 'sentence', 'guilders', 'harshly', 'country', 'rectory'
Topic 5	'dollar', 'included', 'imposed', 'brought', 'theodore', 'analysts', 'recent', 'savings', 'raymond', 'crisis'

Table 6.13: Perplexity comparison for different topic model approaches on Associated Press dataset

K	5	10	15	20	25
NMF	-48927846.38	-14675674.41	-19570125.55	-29349042.83	-34253601.97
LDA	-48914501.38	-14674350.41	-19565800.55	-29348700.83	-34240150.97
CTM	-27347959.58	-4148410.425	-5217722.16	-6290030.703	-7969862.682
MGCTM	-28583395.58	-3715841.425	-5430845.16	-6574180.983	-8003350.762
MGDCTM	-1946680211	-1340688850	-1769709283	-1474757735	-1635640398
MBLCTM	-964784949.7	-1108948218	-1973927828	-1463811648	-1541438023

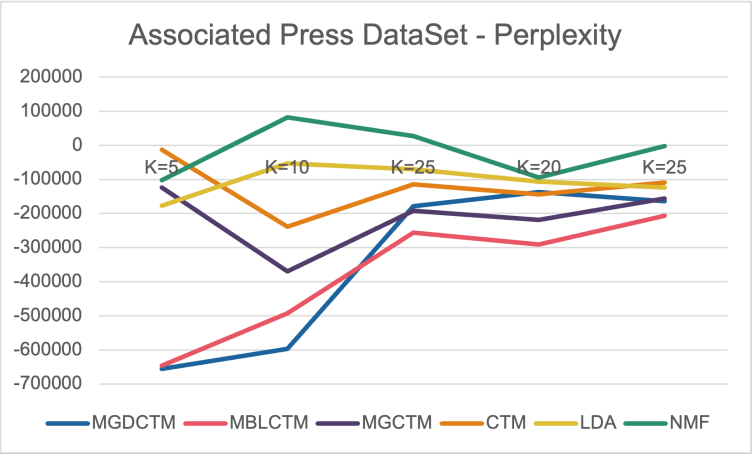


Figure 6.6: Perplexity for Associated Press dataset

Table 6.14: Time complexity comparison for for MPCA, GDMPCA, and BLMPCA at varying topic levels (K) on Associated Press dataset (min)

K	5	10	15	20	25
NMF	19.24	21.164	22.3184	29.2448	30.3992
LDA	18.46	22.152	25.844	29.536	31.382
CTM	25.55	26.01	33.813	35.1135	36.1539
MGCTM	25.32	32.916	35.448	37.98	40.0056
MGDCTM	25.2945	28.105	36.5365	37.94175	39.06595
MBLCTM	25.7259	29.57	39.0324	40.2152	41.398

Chapter 7

Conclusion and Future Work

In this thesis, we proposed and evaluated several novel models to enhance multi-topic modeling and text classification, addressing key limitations of traditional methods. Our research introduced the Generalized Dirichlet Multinomial PCA (GDMPCA) and Beta-Liouville Multinomial PCA (BLMPCA) models, utilizing Bayesian analysis with generalized Dirichlet and Beta-Liouville assumptions. These models demonstrated increased flexibility and superior performance, particularly in text classification and sentiment analysis, as evidenced by higher prediction accuracy compared to the MPCA model. Notably, the BLMPCA showed the most significant improvements across various datasets.

Additionally, we enhanced the bi-RATM method by integrating flexible GD and BL priors. This approach addressed the limitations of previous extensions, such as incomplete generative processes and the use of Dirichlet priors. Applied to text classification and medical text topic modeling, our models achieved superior results with lower perplexity, outperforming established approaches. The BL-bi-RATM, in particular, exhibited remarkable improvements.

Our study further extended the DMR and dDMR models by incorporating GD and BL distributions. This extension effectively handled complex data structures and mitigated overfitting. The use of collapsed Gibbs sampling for parameter inference enhanced the computational capabilities of our models, enabling them to discover latent topics more efficiently. The GD and BL multinomial regression (GDBLMR) models outperformed competing approches, especially with complex and

sparse data, demonstrating improved interpretability and effectiveness in high-dimensional document feature fitting.

Moreover, we integrated GD and BL distributions with Bert-Topic models into the Multi-Grain Clustering Topic Model (MGCTM). The resulting Multi-Grain Generalized Dirichlet Bert-Topic Model (MGGDBTM) and Multi-Grain Beta-Liouville Bert-Topic Model (MGBLBTM) outperformed traditional models like LDA and NMF. These enhanced models better fit high-dimensional document features, reduced overfitting, and improved topic coherence, making them suitable for various applications in natural language processing and machine learning.

Future research could be devoted to modeling modifications and enhancements for greater precision in topic modeling. Potential directions include exploring additional probabilistic distributions, applying models to diverse domains, and incorporating temporal dynamics to track topic evolution. Enhancing scalability and efficiency, improving user interpretability, and integrating with neural networks are crucial next steps. Developing real-time topic modeling systems and leveraging large language models (LLMs) like GPT-3 or GPT-4 to cluster local topics and identify overarching global topics will further enhance topic coherence and relevance. This comprehensive approach paves the way for significant advancements in topic modeling, ensuring its applicability to a wide range of data and real-time streaming scenarios.

Chapter 8

Appendix

8.1 Exponential Family Distribution

The following introduces the general exponential family of distributions:
We have a collection of T functions $t(x)$ and d parameters θ for each specific sample point, represented as a vector of measurements x. These vectors have a size of T and are likely to be influenced by certain additional limitations. The probability distribution $q(x|\theta)$ is as follows [165]:

$$q(x|\theta) = \frac{1}{Y_t(x)Z_t(\theta)} exp(t(x)^{\mp}\theta) \tag{309}$$

$Z_t(\theta)$ is modified to Z, or a distinguishing subscript is inserted. When y is distributed as $q(y|\varrho)$, the notation $E_{q(y|\varrho)}$ is used to describe the expected value of the quantity A. There are two main concepts that must be given [166]:

$$\begin{aligned} \mu_t &\equiv E_{q(y|\varrho)}\{t(x)\} = \frac{\partial \log Z_t}{\partial \theta} \\ \Sigma_t &\equiv E_{q(y|\varrho)}\{(t(x)-\mu_t)(t(x)-\mu_t)^{\mp})\} = \frac{\partial^2 \log Z_t}{\partial\theta\partial\theta} = \frac{\partial \mu_t}{\partial \theta} \end{aligned} \tag{310}$$

The average matrix μ_t has the same number of elements as θ, and the matrix Σ_t represents the covariance of $t(x)$ as mentioned in the reference [19]. Significantly, the variable μ_t acts as

a corresponding element to the parameter set θ. More precisely, when μ_t is completely ordered, it serves as the Hessian for variations in basis. Furthermore, μ_t denotes the anticipated Fisher Information of the distribution. Both t and Σ_t can be obtained directly from Z_t, demonstrating a distinct relationship where μ_t serves as a complimentary parameter set to θ. In situations where μ_t possesses maximum rank, it is instrumental in basis transformations and also signifies the intended Fisher Information for the distribution.

We further detail the characteristics of the exponential family for the Dirichlet, GD, and BL distributions in Table 8.1. Another crucial characteristic of the exponential family is the calculation of maximum a posteriori (MAP) estimations for parameters, obtained from a dataset of I observations. This setup often reflects the structure of a conjugate prior, facilitating the estimation process. One common approach involves the use of an "effective" prior sample size, characterized by relevant statistics ν_t and a prior sample size of S_t. This special method for calculating MAP for parameters within the exponential family provides an approximation for their dual aspects, as explored in [19].

$$\hat{\mu_t} = \frac{\nu_t + \Sigma_i t(x_i)}{S_t + I} \tag{311}$$

Table 8.1: Exponential Family Characterizations for Dirichlet, GD and BL Distributions.

	Characterizations			
MODEL	Z_t	$t_k(x)$	θ_k	$\mu_{t,k}$
Dirichlet	$\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$	$\log(x_1), ..., \log(x_{k+1})$	α_k	$\Psi_0(\alpha_k) - \Psi_($ $\sum_k(\alpha_k))$
GD	$\frac{\Gamma(a_i)\Gamma(b_i)}{\Gamma(a_i+b_i)}$	$\log(x_1), ..., \log(1 - \sum_{t=1}^{D} x_t)$ $- \log(1 - \sum_{t=1}^{D-1} x_t)$	a_k, b_k	$\Big(\Psi(a_i) - \Psi(a_i + b_i) +$ $\sum_{m=1}^{i-1}(\Psi(b_m) - \Psi(a_m + b_m)\Big)$
BL	$\frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$	$\log(x_1) - \log(\sum_{d=1}^{D} x_d), ..., \log x_D$ $- \log(\sum_{d=1}^{D} x_d)$	α_k, α, β	$\Psi(\alpha) - \Psi(\alpha + \beta) +$ $\Psi(\alpha_d) - \Psi(\sum_d \alpha_d)$

8.2 Parameters for GDMPCA

Breaking down of the L parameter for GDMPCA: By factorizing,

$\log p(w|\xi,\Omega) \geq E_q[(\theta,z,w)|\xi,\Omega] - E_q[\log q(z,\theta)]$,

we have 25

In the following, we will derive each of the five factors of the above equation:

$$\begin{aligned} E_q[\log p(\theta|\xi)] = &\sum_{l=1}^{d}[\log\eta(\alpha_l+\beta_l) - \log\eta(\alpha_l) - \log\eta(\beta_l)] \\ &+ \sum_{l=1}^{d}[\alpha_l(\Psi(\eta_l) - \Psi(\eta_l+\vartheta_l)) \\ &+ (\Psi(\vartheta_l) - \Psi(\eta_l+\vartheta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})] \end{aligned} \tag{312}$$

$$\begin{aligned} E_q[\log p(z|\theta)] = &\sum_{n=1}^{N}\sum_{l=1}^{d}\varrho_{nl}(\Psi(\eta_l) - \Psi(\eta_l+\vartheta_l)) \\ &+ \sum_{n=1}^{N}\varrho_{n(d+1)}(\Psi(\vartheta_d) - \Psi(\vartheta_d+\eta_d)) \end{aligned} \tag{313}$$

$$E_q[\log p(w|z,\Omega)] = \sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{v}\varrho_{nl}w_n^j\log(\Omega_{(lj)}) \tag{314}$$

we should mention that $\Omega_{(lj)} = p(w_n^j = 1|z^l = 1)$

$$\begin{aligned} E_q[\log q(\theta)] = &\sum_{l=1}^{d}(\log\eta(\eta_l+\vartheta_l)\log\eta(\eta_l) - \log\eta(\vartheta_l)) \\ &+ \sum_{l=1}^{d}[\eta_l(\Psi(\eta_l) - \Psi(\eta_l+\vartheta_l)) + (\Psi(\vartheta_l) - \Psi(\vartheta_l+\eta_l)) \\ &(\vartheta_l - \eta_{l+1} - \vartheta_{l+1})] \end{aligned} \tag{315}$$

$$E_q[\log q(z)] = \sum_{n=1}^{N}\sum_{l=1}^{D+1}\varrho_{nl}\log(\varrho_{nl}) \tag{316}$$

Next, we will provide a more detailed explanation of Equation 25 by extending it in terms of

both the model characteristics and the variational variables.

$$
\begin{aligned}
\mathcal{L}(\eta, \Phi; \xi, \Omega) = & \sum_{l=1}^{d} [\log \eta(a_l + b_l) - \log \eta(a_l) - \log \eta(b_l)] \\
& + \sum_{l=1}^{d} [a_l(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) \\
& + (\Psi(\Phi) - \Psi(\eta_l + \Phi))(a_l - a_{l+1} - b_{l+1})] \\
& + \sum_{n=1}^{N} \sum_{l=1}^{d} m_{nl}(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + \\
& \sum_{n=1}^{N} m_{n(d+1)}(\Psi(\Phi) - \Psi(\Phi + \eta_d)) \\
& + \sum_{n=1}^{N} \sum_{l=1}^{d+1} \sum_{j=1}^{v} m_{nl} w_n^j \log(\Omega_{ij}) \\
& - \sum_{l=1}^{d} (\log \eta(\eta_l + \Phi) \log \eta(\eta_l) - \log \eta(\Phi)) \\
& - \sum_{l=1}^{d} [\eta_l(\Psi(\eta_l) - \Psi(\eta_l + \Phi)) + (\Psi(\Phi) - \Psi(\Phi + \eta_l)) \\
& (\Phi - \eta_{l+1} - \Phi_{l+1})]
\end{aligned}
\tag{317}
$$

8.2.1 Variational generalized Dirichlet

To obtain revised formulas for the variational parameters in the GD model, you start by isolating the terms in equation 25 that contain the variational parameters of the generalized Dirichlet. This involves examining the equation to identify which parts specifically involve these parameters, then focusing on manipulating these parts to derive expressions for updating the parameters during the variational inference process. This method allows for iterative refinement of the parameters,

enhancing model accuracy with respect to the data being analyzed.

$$
\begin{aligned}
L[\xi_q] = & \sum_{l=1}^{d} [\alpha_l(\Psi(\eta_l) - \Psi(\eta_l + \vartheta_l)) \\
& + (\Psi(\eta_l) - \Psi(\eta_l + \vartheta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})] \\
& + \sum_{n=1}^{N} \varrho_{nl}(\Psi(\eta_l) - \Psi(\eta_l + \vartheta_l) + \sum_{n=1}^{N} \varrho_{n(d+1)}(\Psi(\eta_d) - \Psi(\eta_d + \vartheta_d)) \\
& - \sum_{l=1}^{d} (\log \eta(\eta_l + \vartheta_l) - \log \eta(\eta_l) - \log \eta(\vartheta_l)) \\
& + \sum_{l=1}^{d} (\Psi(\eta_l) - \eta_l(\Psi(\eta_l + \vartheta_l)) \\
& + (\Psi(\vartheta_l) - \Psi(\vartheta_l + \eta_l))(\vartheta_l - \eta_{l+1} - \vartheta_{l+1})))
\end{aligned}
\tag{318}
$$

By approximating the derivative of the aforementioned equation to zero, we obtain the subsequent modified parameters:

$$
\eta_l = \alpha_l + \sum_{n=1}^{N} \varrho_{nl} \tag{319}
$$

$$
\eta_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \varrho_{n(ll)} \tag{320}
$$

Topic based model

To derive the update equations for β_w, maximize equation 25 with respect to β_w. This involves setting the derivatives to zero, mirroring the optimization process used in MPCA, resulting in similar equations.

$$
L[\beta_w] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{K+1} \sum_{j=1}^{V} \varrho_{dnl} w_{dn}^{j} \log \beta_{w(lj)} + \sum_{l=1}^{K+1} \tau_l \big(\sum_{j=1}^{V} \beta_{w(ij)} \big) \tag{321}
$$

By computing the derivative with regard to $\beta_{w(lj)}$ and equating it to zero, we obtain:

$$
\beta_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varrho_{dnl} w_{dn}^{j} \tag{322}
$$

In this scenario, because there are latent variables present in the primary objective function, the

situation isn't fully addressed by Eqs. 35 and 36. However, the probability distribution $q(w|\eta, r, m)$ can be accurately modeled using multinomials, which ensures that the minimum Kullback-Leibler (KL) divergence reaches zero. Consequently, the iterative updates will converge towards a local extremum of the log probability $\log p(\Omega, m|r)$.

$$\eta_l = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} \Omega m_{nl} \tag{323}$$

$$m_{nl} = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} \Omega_{lv} e^{(\tau_n - 1)} e^{(\Psi(\eta_l) - \Psi(\eta_l + \Phi))} \tag{324}$$

$$\Omega_{ij} = \frac{\eta(a_i + b_i)}{\eta(a_i)\eta(b_i)} (2f_j + (\sum_n e^{(\tau_n - 1)} e^{(\Psi(\eta_l) - \Psi(\eta_l + \Phi))}) \tag{325}$$

$$e^{\tau_n - 1} = \frac{1}{\sum_{l=1}^{d} m_{nl} e^{(\Psi(\eta_l) - \Psi(\eta_l + \Phi_l))} + m_{(d+1)n} e^{(\Psi(\Phi_d) - \Psi(\Phi_d + \eta_d))}} \tag{326}$$

Generalized Dirichlet Parameter

We select the components of equation 25 that involve the GD parameters ξ.

$$\begin{aligned} L[\xi] = &\sum_{m=1}^{M} (\log(\eta(\alpha_l + \beta_l)) - \log \eta(\alpha_l)) - \log(\eta(\beta_l))) \\ &+ \sum_{m=1}^{M} (\alpha_l (\Psi(\eta_{ml} - \Psi(\eta_{ml} + \vartheta_{ml})) + \beta_l (\Psi(\vartheta_{ml}) - \Psi(\vartheta_{ml} - \eta_{ml}))) \end{aligned} \tag{327}$$

By differentiating the given equation with respect to the GD parameters, we obtain:

$$\frac{\partial L[\xi]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M} (\Psi(\eta_{ml}) - \Psi(\eta_{ml} + \vartheta_{ml})) \tag{328}$$

and

$$\frac{\partial L[\xi]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M} (\Psi(\vartheta_{ml}) - \Psi(\eta_{ml} + \vartheta_{ml})) \tag{329}$$

The Hessian matrix of the likelihood function in this case assumes a particularly interesting form, as detailed below:

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\alpha_l)] \tag{330}$$

$$\frac{\partial^2 L[\xi]}{\partial \beta_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\beta_l)] \tag{331}$$

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l \beta_l} = M[\Psi^{'}(\alpha_l + \beta_l)] \tag{332}$$

The non-diagonal entries of the Hessian matrix are zero, which imparts a block diagonal structure to the matrix. This configuration simplifies the calculation of the inverse Hessian matrix, as it reduces to inverting the matrices along the diagonal. This simplification allows for an easier derivation of the inverse.

8.3 Variational BLMPCA

In order to calculate the parameter ϱ, which indicates the likelihood that the n-th word is produced by the l-th hidden topic, we optimize the corresponding function by maximizing it with regard to ϱ. This entails fine-tuning the parameter ϱ in order to maximize the probability of the observed data, taking into account the model's assumptions regarding topic distributions.

$$\begin{aligned} L[\varrho_{nl}] = \varrho_{ni}(\Psi(\eta_i) - \Psi(\sum_{l=1}^{D} \eta_l)) + \varrho_{ni} \log \iota_{w(iv)} - \varrho_{ni} \log \varrho_{ni} \\ + \tau_n(\sum_{l=1}^{D} \varrho_{n(l)} - 1) \end{aligned} \tag{333}$$

and

$$\begin{aligned} L[\varrho_{n(D+1)}] = \varrho_{n(D+1)}(\Psi(\iota_\eta - \Psi(\kappa_\eta + \iota_\eta))) + \varrho_{n(D+1)} \log \iota_{(D+1)v} \\ - \varrho_{n(D+1)} \log \varrho_{n(D+1)} + \tau_n(\sum_{i=1}^{D} \varrho_{n(i)} - 1) \end{aligned} \tag{334}$$

Consequently, we have:

$$\frac{\partial L}{\partial \varrho_{nl}} = (\Psi(\eta_d) - \Psi(\sum_{l=1}^{D} \eta_l)) + \log \iota_{w(iv)} - \log \varrho_{ni} - 1 + \tau_n \tag{335}$$

and

$$\frac{\partial L}{\partial \varrho_{n(D+1)}} = (\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta)) \tag{336}$$

Equating the preceding equation to zero yields:

$$\varrho_{nl} = \iota_{lv} e^{(\tau_n - 1)} e^{(\Psi(\eta_i) - \Psi(\sum_{ii=1}^{D} \eta_{ii}))} \tag{337}$$

$$\varrho_{n(D+1)} = \iota_{(D+1)v} e^{(\tau_n - 1)} e^{(\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta))} \tag{338}$$

considering that $\sum_{d=1}^{D+1} \varrho_{n(d)} = 1$ for the normalization factor we have:

$$e^{\tau_n - 1} = \frac{1}{\iota_{(D+1)v} e^{(\tau_n - 1)} e^{(\Psi(\iota_\eta) - \Psi(\kappa_\eta + \iota_\eta))} + \iota_{lv} e^{(\tau_n - 1)} e^{(\Psi(\eta_i) - \Psi(\sum_{ii=1}^{D} \eta_{ii}))}} \tag{339}$$

8.3.1 Variational Beta-Liouville

The updates mentioned are designed to converge to a local maximum of a lower bound of $\log p(\Omega, \Upsilon | r)$, which is optimal for all product approximations such as $q(m)q(w)$ for the joint probability $p(m, w|\Omega, \Upsilon, r)$. This approach ensures that the variational parameters are fine-tuned to best approximate the true posterior distributions within the constraints of the model.

$$\Phi_l = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D} \kappa_d)\eta(\kappa + \iota)} m_{nl}(\tau_n - 1)(\Psi(\eta_l) - \Psi(\sum_{l=1}^{D} \eta_l) \tag{340}$$

$$\eta_l = \kappa_l + \sum_{n=1}^{N} m_{nl} \tag{341}$$

$$\Omega_{(lj)} = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D} \kappa_d)\eta(\kappa + \iota)} (2f \sum_{d=1}^{M} \sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j}) \tag{342}$$

In this case, variable Ω vanishes because m is defined in terms of the KL approximation. In the second step, the algorithm now optimizes for m. Since $q(w|\eta, r, m)$ can be precisely modelled with multinomials, the minimum KL divergence is zero. As a result, the updates that follow converge to

a local threshold of $\log p(\Omega, m|r)$.

$$\eta_l = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D}\kappa_d)\eta(\kappa+\iota)}\Omega m_{nl} \tag{343}$$

$$m_{nl} = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D}\kappa_d)\eta(\kappa+\iota)}\Omega_{lv}e^{(\tau_n-1)}e^{(\Psi(\eta_i)-\Psi(\sum_{ii=1}^{D}\eta_{ii})} \tag{344}$$

$$\Omega_{ij} = \frac{\eta(\kappa)\eta(\iota)}{\eta(\sum_{d=1}^{D}\kappa_d)\eta(\kappa+\iota)}(2f + (\sum_{n} e^{(\tau_n-1)}e^{(\Psi(\eta_i)-\Psi(\sum_{ii=1}^{D}\eta_{ii})}) \tag{345}$$

considering that $\sum_{d=1}^{D+1}\varrho_{n(d)} = 1$ for the normalization factor we have :

$$e^{\tau_n-1} = \frac{1}{m_{(D+1)v}e^{(\tau_n-1)}e^{(\Psi(\iota_\eta)-\Psi(\kappa_\eta+\iota_\eta))} + m_{lv}e^{(\tau_n-1)}e^{(\Psi(\eta_i)-\Psi(\sum_{ii=1}^{D}\eta_{ii}))}} \tag{346}$$

8.4 Parameters for GDMPCA

Breaking down of the L parameter for GDMPCA:

We can factor, $\log p(w|\xi,\beta_w) \geq E_q[(\theta,z,w)|\xi,\beta_w] - E_q[\log q(z,\theta)]$ to obtain the equation:

$$\begin{aligned} L(\xi_q,\Phi_w;\xi,\beta_w) = {} & E_q[\log p(\theta|\xi)] + E_q[\log p(z)] + E_q[\log p(w|z,\beta_w)] \\ & - E_q[\log q(\theta)] - E_q[\log q(z)] \end{aligned} \tag{347}$$

The subsequent section will elaborate on how each of the five components in the previous equation is obtained:

$$\begin{aligned} E_q[\log p(\theta|\xi)] = {} & \sum_{l=1}^{d}[\log\Gamma(\alpha_l+\beta_l) - \log\Gamma(\alpha_l) - \log\Gamma(\beta_l)] \\ & + \sum_{l=1}^{d}[\alpha_l(\Psi(\gamma_l) - \Psi(\gamma_l+\delta_l)) \\ & + (\Psi(\delta_l) - \Psi(\gamma_l+\delta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})] \end{aligned} \tag{348}$$

$$E_q[\log p(z|\theta)] = \sum_{n=1}^{N}\sum_{l=1}^{d}\phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \sum_{n=1}^{N}\phi_{n(d+1)}(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d)) \tag{349}$$

$$E_q[\log p(w|z, \beta_w)] = \sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{v}\phi_{nl}w_n^j \log(\beta_{w(lj)}) \tag{350}$$

we should mention that $\beta_{w(lj)} = p(w_n^j = 1|z^l = 1)$

$$\begin{aligned} E_q[\log q(\theta)] = &\sum_{l=1}^{d}(\log\Gamma(\gamma_l + \delta_l)\log\Gamma(\gamma_l) - \log\Gamma(\delta_l)) \\ &+ \sum_{l=1}^{d}[\gamma_l(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + (\Psi(\delta_l) - \Psi(\delta_l + \gamma_l)) \\ &(\delta_l - \gamma_{l+1} - \delta_{l+1})] \end{aligned} \tag{351}$$

$$E_q[\log q(z)] = \sum_{n=1}^{N}\sum_{l=1}^{D+1}\phi_{nl}\log(\phi_{nl}) \tag{352}$$

8.4.1 Varitional Multinomial

In order to find ϕ_{nl} we proceed to maximize with the respect to ϕ_{nl} so we have following equations:

$$\begin{aligned} L[\phi_{nl}] = &\phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \phi_{nl}\log\beta_{w(lv)} - \phi_{nl}\log\phi_{nl} \\ &+ \lambda_n(\sum_{ll=1}^{d+1}\phi_{n(ll)} - 1) \end{aligned} \tag{353}$$

and

$$\begin{aligned} L[\phi_{n(d+1)}] = &\phi_{n(d+1)}(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d) + \phi_{n(D+1)}\log\beta_{(d+1)v} \\ &- \phi_{n(d+1)}\log\phi_{n(d+1)} + \lambda_n(\sum_{ll=1}^{d+1}\phi_{n(ll)} - 1) \end{aligned} \tag{354}$$

and therefore we have:

$$\frac{\partial L}{\partial \phi_{nl}} = (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \log \beta_{lv} - \log \phi_{nl} - 1 + \lambda_n \tag{355}$$

and

$$\frac{\partial L}{\partial \phi_{n(d+1)}} = (\Psi(\gamma_d) - \Psi(\gamma_d + \delta_d)) + \log \beta_{(d+1)v} - \log \phi_{n(d+1)} - 1 + \lambda_n \tag{356}$$

the equation can be solved by setting it to zero, which yields the value of:

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} \tag{357}$$

$$\phi_{n(d+1)} = \beta_{(d+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d))} \tag{358}$$

To ensure normalization, we have $\sum_{ll=1}^{d+1} \phi_{n(ll)} = 1$, where $n(ll)$ represents the ll-th word position in the document, we have :

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} \beta_{lv} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} + \beta_{(d+1)v} e^{(\Psi(\delta_d) - \Psi(\delta_d + \gamma_d))}} \tag{359}$$

8.4.2 Variational generalized Dirichlet

In order to obtain the update equations for the variational generalized Dirichlet, we will separate the terms that involve the variational generalized Dirichlet parameters in equation 347.

$$\begin{aligned} L[\xi_q] = & \sum_{l=1}^{d} [\alpha_l (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + (\Psi(\gamma_l) - \\ & \Psi(\gamma_l + \delta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})] \\ & + \sum_{n=1}^{N} \phi_{nl} (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l) + \sum_{n=1}^{N} \phi_{n(d+1)} (\Psi(\gamma_d) - \Psi(\gamma_d + \delta_d)) \\ & - \sum_{l=1}^{d} (\log \Gamma(\gamma_l + \delta_l) - \log \Gamma(\gamma_l) - \log \Gamma(\delta_l)) + \sum_{l=1}^{d} (\Psi(\gamma_l) - \\ & \gamma_l (\Psi(\gamma_l + \delta_l)) + (\Psi(\delta_l) - \Psi(\delta_l + \gamma_l))(\delta_l - \gamma_{l+1} - \delta_{l+1}))) \end{aligned} \tag{360}$$

to obtain the updated parameters, we can set the derivative of the above equation to zero:

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \tag{361}$$

$$\gamma_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \tag{362}$$

Topic based model

To estimate the parameter β_w, we obtain the update equations by maximizing the equation 347. This process leads to the same set of equations as that of MPCA. The resulting equations are:

$$L[\beta_w] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{K+1} \sum_{j=1}^{V} \phi_{dnl} w_{dn}^{j} \log \beta_{w(lj)} + \sum_{l=1}^{K+1} \lambda_l \big(\sum_{j=1}^{V} \beta_{w(ij)} \big) \tag{363}$$

To obtain the optimal value of $\beta_{w(lj)}$, we differentiate the equation with respect to $\beta_{w(lj)}$ and equate it to zero, resulting in:

$$\beta_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \tag{364}$$

Generalized Dirichlet Parameter

We select the expressions of the equation 347 that include the generalized Dirichlet parameters ξ.

$$\begin{aligned} L[\xi] = & \sum_{m=1}^{M} (\log(\Gamma(\alpha_l + \beta_l)) - \log \Gamma(\alpha_l)) - \log(\Gamma(\beta_l))) \\ & + \sum_{m=1}^{M} (\alpha_l(\Psi(\gamma_{ml} - \Psi(\gamma_{ml} + \delta_{ml})) + \beta_l(\Psi(\delta_{ml}) - \Psi(\delta_{ml} - \gamma_{ml}))) \end{aligned} \tag{365}$$

To calculate the derivative of the given equation with respect to the generalized Dirichlet parameters, we have:

$$\frac{\partial L[\xi]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M} (\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{366}$$

and

$$\frac{\partial L[\xi]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M}(\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{367}$$

The Newton-Raphson method is utilized to solve the equation, and the Hessian matrix in respect to the parameter space is required for this purpose. Interestingly, the Hessian matrix of the likelihood takes a peculiar form which is as follows:

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\alpha_l)] \tag{368}$$

$$\frac{\partial^2 L[\xi]}{\partial \beta_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\beta_l)] \tag{369}$$

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l \beta_l} = M[\Psi^{'}(\alpha_l + \beta_l)] \tag{370}$$

The Hessian matrix has only non-zero entries on its diagonal, as shown in the above equations. This results in the Hessian matrix having a block diagonal form, and therefore the inverse Hessian matrix can be derived easily as the inverse of the matrix on the diagonal.

8.5 Variational Bete-Louisville distribution

To obtain the parameter representing the probability that the nth word is generated by the l-th hidden topic, we aim to maximize the expression with respect to ϕ:

$$\begin{aligned} L[\phi_{nl}] = \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l)) + \phi_{ni}\log\beta_{w(iv)} \\ - \phi_{ni}\log\phi_{ni} + \lambda_n(\sum_{l=1}^{D}\phi_{n(l)} - 1) \end{aligned} \tag{371}$$

and

$$\begin{aligned} L[\phi_{n(D+1)}] = \phi_{n(D+1)}(\Psi(\beta_\gamma - \Psi(\alpha_\gamma + \beta_\gamma))) + \phi_{n(D+1)}\log\beta_{(D+1)v} \\ - \phi_{n(D+1)}\log\phi_{n(D+1)} + \lambda_n(\sum_{i=1}^{D}\phi_{n(i)} - 1) \end{aligned} \tag{372}$$

and therefore we have:

$$\frac{\partial L}{\partial \phi_{nl}} = (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \log \beta_{w(iv)} - \log \phi_{ni} - 1 + \lambda_n \tag{373}$$

and

$$\frac{\partial L}{\partial \phi_{n(D+1)}} = (\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \tag{374}$$

setting the above equation to zero leads to:

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))} \tag{375}$$

$$\phi_{n(D+1)} = \beta_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} \tag{376}$$

Given that the sum of $\phi_{n(d)}$ from $d = 1$ to $D + 1$ is equal to 1, which serves as the normalization factor, we can conclude that:

$$e^{\lambda_n - 1} = \frac{1}{\beta_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} + \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))}} \tag{377}$$

8.5.1 Variational Beta-Liouville

In order to derive the update equations for the variational BL, we will once again follow the process of seprating the terms that involve the variation BL parameters.

$$\begin{aligned}
L[\xi_q] = & \alpha_d(\Psi(\gamma_d)) - \Psi(\sum_{l=1}^{D}\gamma_l) + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \beta(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N}\phi_{n(D+1)}(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& - (\log(\Gamma(\sum_{l=1}^{D}\gamma_l)) + \log(\gamma(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) \\
& - \log(\Gamma(\beta_\gamma)) - \log(\Gamma(\gamma_i)) \\
& + \gamma_i(\Psi(\gamma_i) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& - \Psi(\sum_{l=1}^{D}\gamma_l) + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \beta_\gamma(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)))
\end{aligned} \tag{378}$$

By choosing the expressions that involve the variational BL parameters γ_i, α_γ, and β_γ, we can obtain:

$$\begin{aligned}
L(\gamma_i) = & \alpha_i(\Psi(\gamma_i)) - (\sum_{l=1}^{D}\alpha_l)(\Psi(\sum_{l=1}^{D}\gamma_l)) + \sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l)) \\
& - (\log\Gamma(\sum_{l=1}^{D}) - \log\Gamma(\gamma_i) + \gamma_i(\Psi(\sum_{l=1}^{D}\gamma_l)\sum_{d=1}^{D}\gamma_d)
\end{aligned} \tag{379}$$

and

$$
\begin{aligned}
L[\alpha_\gamma] = & \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(-\Psi(\alpha_\gamma + \beta_\gamma)) \\
& + (\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \sum_{n=1}^{N} \sum_{i=1}^{D} \phi_{ni} \\
& + \sum_{n=1}^{N} \phi_{n(D+1)}(-\Psi(\alpha_\gamma + \beta_\gamma)) - (\log(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) + \\
& \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta_\gamma(-\Psi(\alpha_\gamma + \beta_\gamma)))
\end{aligned}
\tag{380}
$$

If we differentiate the aforementioned equations with respect to their corresponding BL parameters, we obtain:

$$
\begin{aligned}
\frac{\partial L[\gamma_i]}{\partial \gamma_i} = & \alpha_i \Psi^{'}(\gamma_i) - \Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{l=1}^{D} \alpha_l + \Psi^{'}(\gamma_i) \sum_{n=1}^{N} \phi_{ni} \\
& - D\Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{n=1}^{N} \phi_{ni} - (\Psi(\sum_{l=1}^{D} \gamma_l) + \gamma_i \Psi^{'}(\gamma_i) \\
& - \Psi^{'}(\sum_{l=1}^{D} \gamma_l) \sum_{d=1}^{D} \gamma_l - \psi(\sum_{l=1}^{D} \gamma_l))
\end{aligned}
\tag{381}
$$

and

$$
\begin{aligned}
\frac{\partial L[\gamma_i]}{\partial \alpha_\gamma} = & \alpha(\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) - \beta(\Psi^{'}(\alpha_\gamma + \beta_\gamma)) \\
& + (\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) \sum_{n=1}^{N} \sum_{d=1}^{D} \phi_{nd} \\
& - \sum_{n=1}^{N} \phi_{n(D+1)}(\Psi^{'}(\alpha_\gamma + \beta_\gamma)) - (\alpha_\gamma(\Psi^{'}(\alpha_\gamma) - \Psi^{'}(\alpha_\gamma + \beta_\gamma)) \\
& - \beta_\gamma(\Psi^{'}(\alpha_\gamma + \beta_\gamma))
\end{aligned}
\tag{382}
$$

The update equations for the variational BL can be obtained by equating the previously derived equations to zero.

$$
\gamma_i = \alpha + \sum_{n=1}^{N} \phi_{ni} \tag{383}
$$

$$\alpha_\gamma = \alpha + \sum_{n=1}^{N}\sum_{d=1}^{D}\phi_{nd} \tag{384}$$

$$\beta_\gamma = \beta + \sum_{n=1}^{N}\phi_{n(D+1)} \tag{385}$$

Topic Based Multinomial

In this section, we will obtain the required update equations for estimating β_w. When we maximize the equation 347 with respect to β_w, we arrive at the same equation as in the MPCA scenario.

$$L[\beta_w] = \sum_{d=1}^{M}\sum_{n=1}^{N_s}\sum_{l=1}^{D+1}\sum_{j=1}^{V}\phi_{dnl}w_{dn}^{j}\log\beta_{w(lj)} + \sum_{l=1}^{D+1}\lambda_l(\sum_{j=1}^{V}\beta_{w(lj)} - 1) \tag{386}$$

By differentiating with respect to $\beta_{w(lj)}$ and equating to zero, we obtain:

$$\beta_{w(lj)} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dnl}w_{dn}^{j} \tag{387}$$

Beta-Liouville Parameters

$$\begin{aligned} L[\xi] = &\sum_{m=1}^{M}(\log(\Gamma(\sum_{l=1}^{D}\alpha_l)) + \log(\Gamma(\alpha+\beta)) - \log\Gamma(\alpha) \\ &- \log\Gamma(\beta) - \sum_{i=1}^{D}\log\Gamma(\alpha_i) + \sum_{i=1}^{D}\alpha_i(\Psi(\gamma_{mi}) \\ &- \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) + \alpha(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}\beta_{m\gamma})) + \beta(\Psi(\beta_{m\gamma}) \\ &- \Psi(\alpha_{m\gamma}+\beta_{m\gamma}))) \end{aligned} \tag{388}$$

The expression for the derivative of the above equation with respect to the BL parameter can be expressed as:

$$\begin{aligned}
\frac{\partial L[\xi]}{\partial \alpha_l} &= M(\Psi(\sum_{l=1}^{D}) - \Psi(\alpha_l)) + \sum_{m=1}^{M}(\Psi^{'}(\gamma_{ml}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)})) \\
\frac{\partial L[\xi]}{\partial \alpha} &= M[\Psi(\alpha+\beta) - \Psi(\alpha)] + \sum_{m=1}^{M}(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma})) \\
\frac{\partial L[\xi]}{\partial \beta} &= M[\Psi(\alpha+\beta) - \Psi(\beta)] + \sum_{m=1}^{M}(\Psi(\beta_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma}))
\end{aligned} \tag{389}$$

As evident from the aforementioned equations, the derivative of 347 with respect to the BL parameters depends on not only their own values but also on one another. To address this optimization problem, we will employ the Newton-Raphson method. To execute the Newton-Raphson method, it is necessary to calculate the Hessian matrix with respect to the parameter space in the following manner:

$$\begin{aligned}
\frac{\partial^2 L[\xi]}{\partial \alpha_l \alpha_j} &= M(-\delta(i,j)\Psi^{'}(\alpha_i) + \Psi^{'}(\sum_{l=1}^{D}\alpha_l)) \\
\frac{\partial^2 L[\xi]}{\partial \alpha^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\alpha)) \\
\frac{\partial^2 L[\xi]}{\partial \alpha \partial \beta} &= M\Psi^{'}(\alpha+\beta) \\
\frac{\partial^2 L[\xi]}{\partial \beta^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\beta))
\end{aligned} \tag{390}$$

Bibliography

[1] Charu C. Aggarwal. An introduction to cluster analysis. In Charu C. Aggarwal and Chandan K. Reddy, editors, *Data Clustering: Algorithms and Applications*, pages 1–28. CRC Press, 2013.

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

[3] Jianchang Mao and Anil K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Networks*, 6(2):296–317, 1995.

[4] Shipeng Yu, Kai Yu, Volker Tresp, and Hans-Peter Kriegel. A probabilistic clustering-projection model for discrete data. In *European conference on principles of data mining and knowledge discovery*, pages 417–428. Springer, 2005.

[5] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, pages 2267–2273, 2015.

[6] Advaith Siddharthan. Inderjeet mani and mark t. maybury (eds). *Advances in Automatic Text Summarization*. MIT press, 1999. ISBN 0-262-13359-8, 442 pp. *Nat. Lang. Eng.*, 7(3):271–274, 2001.

[7] Doug Beeferman, Adam L. Berger, and John D. Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, 1999.

[8] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multim. Tools Appl.*, 78(11):15169–15211, 2019.

[9] Ronen Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, 2013.

[10] Ting Hua, Chang-Tien Lu, Jaegul Choo, and Chandan K. Reddy. Probabilistic topic modeling for comparative analysis of document collections. *ACM Trans. Knowl. Discov. Data*, 14(2):24:1–24:27, 2020.

[11] David A. Cohn and Thomas Hofmann. The missing link - A probabilistic model of document content and hypertext connectivity. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000*, pages 430–436, 2000.

[12] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[14] Chris Ding, Xiaofeng He, Hongyuan Zha, and Horst D Simon. Adaptive dimension reduction for clustering high dimensional data. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 147–154. IEEE, 2002.

[15] Tao Li, Sheng Ma, and Mitsunori Ogihara. Document clustering via adaptive subspace iteration. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, pages 218–225, 2004.

[16] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[17] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8,2001, Vancouver, British Columbia, Canada]*, pages 617–624, 2001.

[18] Wray Buntine. Variational extensions to em and multinomial pca. In *European Conference on Machine Learning*, pages 23–34. Springer, 2002.

[19] Nicolas Jouvin, Pierre Latouche, Charles Bouveyron, Guillaume Bataillon, and Alain Livartowski. Clustering of count data through a mixture of multinomial pca. *arXiv preprint arXiv:1909.00721*, 2019.

[20] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004.

[21] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems.*, pages 856–864, 2010.

[22] William J. Fitzgerald. Markov chain monte carlo methods with applications to signal processing. *Signal Process.*, 81(1):3–18, 2001.

[23] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.

[24] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[25] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*, 2020.

[26] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[27] Pengtao Xie and Eric P Xing. Integrating document clustering and topic modeling. *arXiv preprint arXiv:1309.6874*, 2013.

[28] Wenxin Liang, Ran Feng, Xinyue Liu, Yuangang Li, and Xianchao Zhang. Gltm: A global and local word embedding-based topic model for short texts. *IEEE access*, 6:43612–43621, 2018.

[29] Mohammad Alhawarat and M Hegazi. Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access*, 6:42740–42749, 2018.

[30] Advaith Siddharthan. Christopher d. manning and hinrich schutze. *Foundations of Statistical Natural Language Processing*. MIT press, 2000. ISBN 0-262-13360-1, 620 pp. $64.95/£44.95 (cloth). *Nat. Lang. Eng.*, 8(1):91–92, 2002.

[31] Simon Lacoste-Julien, Fei Sha, and Michael Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in neural information processing systems*, 21, 2008.

[32] Maxim Rabinovich and David M. Blei. The inverse regression topic model. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 199–207, 2014.

[33] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, a Special Interest Group of the ACL*, pages 248–256, 2009.

[34] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 241–248, 2006.

[35] Tao Ge, Wenzhe Pei, Heng Ji, Sujian Li, Baobao Chang, and Zhifang Sui. Bring you to the past: Automatic generation of topically relevant event chronicles. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 575–585, 2015.

[36] Aytuğ Onan. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, 7:145614–145633, 2019.

[37] Shuangyin Li, Yu Zhang, and Rong Pan. Bi-directional recurrent attentional topic model. *ACM Trans. Knowl. Discov. Data*, 14(6):74:1–74:30, 2020.

[38] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 2717–2725, 2012.

[39] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, pages 978–1, 2012.

[40] Aytug Onan and Mansur Alp Toçoğlu. A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification. *IEEE Access*, 9:7701–7722, 2021.

[41] Gaurav Meena, Krishna Kumar Mohbey, Ajay Indian, Mohammad Zubair Khan, and Sunil Kumar. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimedia Tools and Applications*, pages 1–22, 2023.

[42] Fatma Najar and Nizar Bouguila. Emotion recognition: A smoothed dirichlet multinomial solution. *Eng. Appl. Artif. Intell.*, 107:104542, 2022.

[43] Gaurav Meena, Krishna Kumar Mohbey, and Sunil Kumar. Sentiment analysis on images using convolutional neural networks based inception-v3 transfer learning approach. *International Journal of Information Management Data Insights*, 3(1):100174, 2023.

[44] Gaurav Meena, Krishna Kumar Mohbey, Sunil Kumar, and K Lokesh. A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. *Decision Analytics Journal*, 7:100243, 2023.

[45] 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):391–407, 1990.

[46] Ali Shojaee Bakhtiari and Nizar Bouguila. A variational bayes model for count data learning and classification. *Eng. Appl. Artif. Intell.*, 35:176–186, 2014.

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

[48] Ali Shojaee Bakhtiari and Nizar Bouguila. A latent beta-liouville allocation model. *Expert Syst. Appl.*, 45:260–272, 2016.

[49] David M. Blei and John D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 147–154, 2005.

[50] Lan Du, Wray L. Buntine, Huidong Jin, and Changyou Chen. Sequential latent dirichlet allocation. *Knowl. Inf. Syst.*, 31(3):475–503, 2012.

[51] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[52] Yong Chen, Junjie Wu, Jianying Lin, Rui Liu, Hui Zhang, and Zhiwen Ye. Affinity regularized non-negative matrix factorization for lifelong topic modeling. *IEEE Trans. Knowl. Data Eng.*, 32(7):1249–1262, 2020.

[53] Zhiyuan Chen and Bing Liu. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31th International Conference on Machine Learning, ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 703–711, 2014.

[54] Mingyang Xu, Ruixin Yang, Steve Harenberg, and Nagiza F. Samatova. A lifelong learning topic model structured using latent embeddings. In *11th IEEE International Conference on Semantic Computing, ICSC*, pages 260–261, 2017.

[55] Ilya Sutskever, James Martens, and Geoffrey E. Hinton. Generating text with recurrent neural networks. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1017–1024, 2011.

[56] Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA, 2010.

[57] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111–3119, 2013.

[58] Amr Ahmed and Eric P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the SIAM International Conference on Data Mining, SDM*, pages 219–230, 2008.

[59] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J. Smola, and Le Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–228, 2015.

[60] Zhe Gan, Changyou Chen, Ricardo Henao, David E. Carlson, and Lawrence Carin. Scalable deep poisson factor analysis for topic modeling. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1823–1832, 2015.

[61] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *J. Mach. Learn. Res.*, 17:163:1–163:44, 2016.

[62] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174, 2016.

[63] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Trans. Assoc. Comput. Linguistics*, 3:299–313, 2015.

[64] Koffi Eddy Ihou, Nizar Bouguila, and Wassim Bouachir. Efficient integration of generative topic models into discriminative classifiers using robust probabilistic kernels. *Pattern Anal. Appl.*, 24(1):217–241, 2021.

[65] Samar Hannachi, Fatma Najar, Koffi Eddy Ihou, and Nizar Bouguila. Collapsed gibbs sampling of beta-liouville multinomial for short text clustering. In Hamido Fujita, Ali Selamat, Jerry Chun-Wei Lin, and Moonis Ali, editors, *Advances and Trends in Artificial Intelligence.*

Artificial Intelligence Practices - 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26-29, 2021, Proceedings, Part I, volume 12798 of *Lecture Notes in Computer Science*, pages 564–571. Springer, 2021.

[66] David M. Blei and Jon D. McAuliffe. Supervised topic models. 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 121–128. Curran Associates, Inc., 2007.

[67] David M. Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *CoRR*, abs/1206.3278, 2012.

[68] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2210–2216. AAAI Press, 2015.

[69] Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. End-to-end learning of lda by mirror-descent back propagation over a deep architecture. *Advances in Neural Information Processing Systems*, 28, 2015.

[70] Li Wan, Leo Zhu, and Rob Fergus. A hybrid neural network-latent topic model. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pages 1287–1294. JMLR.org, 2012.

[71] Adrian Benton and Mark Dredze. Deep dirichlet multinomial regression. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 365–374. Association for Computational Linguistics, 2018.

[72] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 452–456. SIAM, 2004.

[73] D Dhanush, Abhinav Kumar Thakur, and Narasimha Prasad Diwakar. Aspect-based sentiment summarization with deep neural networks. *International Journal of Engineering Research & Technology (IJERT)*, 5(5), 2016.

[74] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[75] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[76] Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[77] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 33–42. IEEE, 2017.

[78] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[79] Thomas Minka. Estimating a dirichlet distribution. page 1, 01 2003.

[80] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[81] Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Comput. Linguistics*, 18(4):467–479, 1992.

[82] Volkmar Frinken, Andreas Fischer, R. Manmatha, and Horst Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):211–224, 2012.

[83] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, 2010.

[84] Jonathan D. Chang and David M. Blei. Relational topic models for document networks. In David A. Van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, pages 81–88. JMLR.org, 2009.

[85] Mirwaes Wahabzada, Zhao Xu, and Kristian Kersting. Topic models conditioned on relations. In José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, volume 6323 of *Lecture Notes in Computer Science*, pages 402–417. Springer, 2010.

[86] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

[87] Hali Edison and Hector Carcel. Text data analysis using latent dirichlet allocation: an application to fomc transcripts. *Applied Economics Letters*, 28(1):38–42, 2021.

[88] Zhiwen Luo, Manar Amayri, Wentao Fan, and Nizar Bouguila. Cross-collection latent beta-liouville allocation model training with privacy protection and applications. *Applied Intelligence*, 53(14):17824–17848, 2023.

[89] Fatma Najar and Nizar Bouguila. Sparse document analysis using beta-liouville naive bayes with vocabulary knowledge. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 351–363. Springer, 2021.

[90] 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.

[91] Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Trans. Knowl. Data Eng.*, 20(4):462–474, 2008.

[92] Nizar Bouguila and Djemel Ziou. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1716–1731, 2007.

[93] Pantea Koochemeshkian, Nuha Zamzami, and Nizar Bouguila. Flexible distribution-based regression models for count data: Application to medical diagnosis. *Cybern. Syst.*, 51(4):442–466, 2020.

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

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

[96] Karla L. Caballero Espinosa, Joel Barajas, and Ram Akella. The generalized dirichlet distribution in enhanced topic detection. In Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 773–782. ACM, 2012.

[97] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1353–1360. MIT Press, 2006.

[98] Sahar Salmanzade Yazdi, Fatma Najar, and Nizar Bouguila. Bayesian folding-in using generalized dirichlet and beta-liouville kernels for information retrieval. In *IEEE Symposium Series on Computational Intelligence, SSCI 2022, Singapore, December 4-7, 2022*, pages 1430–1435. IEEE, 2022.

[99] Ali Shojaee Bakhtiari. *Count Data Modeling and Classification Using Statistical Hierarchical Approaches and Multi-topic Models*. PhD thesis, Concordia University, 2014.

[100] Koffi Eddy Ihou and Nizar Bouguila. Stochastic topic models for large scale and nonstationary data. *Engineering Applications of Artificial Intelligence*, 88:103364, 2020.

[101] John Horgan. From complexity to perplexity. *Scientific American*, 272(6):104–109, 1995.

[102] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[103] Ellen Riloff and Wendy Lehnert. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12(3):296–333, 1994.

[104] Hanna M. Wallach. Topic modeling: beyond bag-of-words. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)*, volume 148 of *ACM International Conference Proceeding Series*, pages 977–984, 2006.

[105] David Bamman and Noah A. Smith. New alignment methods for discriminative book summarization. *CoRR*, abs/1305.1319, 2013.

[106] Aytuğ Onan. Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine*, 2018, 2018.

[107] Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14(1):1–15, 2013.

[108] Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53, 2010.

[109] Amir Karami, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. Flatm: A fuzzy logic approach topic model for medical documents. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–6. IEEE, 2015.

[110] Amir Karami, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections. *iConference 2015 Proceedings*, 2015.

[111] Tmvar dataset. available at https://www.ncbi.nlm.nih.gov/research/bionlp/.

[112] Amir Karami, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. Fuzzy approach topic discovery in health and medical corpora. *International Journal of Fuzzy Systems*, 20(4):1334–1345, 2018.

[113] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38, 2011.

[114] Fatma Najar and Nizar Bouguila. Smoothed generalized dirichlet: A novel count-data model for detecting emotional states. *IEEE Trans. Artif. Intell.*, 3(5):685–698, 2022.

[115] Aytuğ Onan. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23):e5909, 2021.

[116] Xiaowei Yan, Guangmin Li, Qian Li, Jiejie Chen, Wenjing Chen, and Fan Xia. sentiment analysis on massive open online course evaluation. In *2021 International Conference on Neuromorphic Computing (ICNC)*, pages 245–249. IEEE, 2021.

[117] Fatma Najar and Nizar Bouguila. Sentiment analysis using smoothed probabilistic-based models. In *9th International Conference on Control, Decision and Information Technologies, CoDIT 2023, Rome, Italy, July 3-6, 2023*, pages 1185–1190. IEEE, 2023.

[118] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, 62:1–16, 2016.

[119] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

[120] Claudia Aparecida Martins, Maria Carolina Monard, and Edson Takashi Matsubara. Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications*, pages 228–233, 2003.

[121] Nuha Zamzami and Nizar Bouguila. Mml-based approach for determining the number of topics in EDCM mixture models. In *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.

[122] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[123] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1422–1432, 2015.

[124] Jader Abreu, Luis Fred, David Macêdo, and Cleber Zanchettin. Hierarchical attentional hybrid neural networks for document classification. In Igor V. Tetko, Vera Kurková, Pavel

Karpov, and Fabian J. Theis, editors, *International Conference on Artificial Neural Networks*, volume 11731 of *Lecture Notes in Computer Science*, pages 396–402, 2019.

[125] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. *Advances in neural information processing systems*, 22:1973–1981, 2009.

[126] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382, pages 1105–1112, 2009.

[127] Nuha Zamzami and Nizar Bouguila. Sparse count data clustering using an exponential approximation to generalized dirichlet multinomial distributions. *IEEE Trans. Neural Networks Learn. Syst.*, 33(1):89–102, 2022.

[128] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[129] Xitong Yang. Understanding the variational lower bound. *Tech. Rep*, 13:1–4, 2017.

[130] Mtranscript dataset. available at https://www.mtsamples.com.

[131] Mentalhealth dataset. available at https://www.thekimfoundation.org/faqs/, https://www.mhanational.org/frequently-asked-questions, https://www.wellnessinmind.org/frequently-asked-questions/, https://www.heretohelp.bc.ca/questions-and-answers.

[132] Genia corpus. available at http://www.geniaproject.org/genia-corpus.

[133] Behrang Mohit. Named entity recognition. *Natural language processing of semitic languages*, pages 221–245, 2014.

[134] Jenn Riley. Understanding metadata. *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)*, 23:7–10, 2017.

[135] Koffi Eddy Ihou, Nizar Bouguila, and Wassim Bouachir. Efficient integration of generative topic models into discriminative classifiers using robust probabilistic kernels. *Pattern Anal. Appl.*, 24(1):217–241, 2021.

[136] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5220–5227, 2004.

[137] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of artificial intelligence research*, 30:249–272, 2007.

[138] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 233–240. ACM, 2007.

[139] Adrian Benton, Michael J. Paul, Braden Hancock, and Mark Dredze. Collective supervision of topic models for predicting surveys with social media. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2892–2898. AAAI Press, 2016.

[140] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM, 2008.

[141] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[142] Ian Porteous, David Newman, Alexander Ihler, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 569–577. ACM, 2008.

[143] Nizar Bouguila and Djemel Ziou. A hybrid sem algorithm for high-dimensional unsupervised learning using a finite generalized dirichlet mixture. *IEEE Transactions on Image Processing*, 15(9):2657–2668, 2006.

[144] Yiwen Zhang, Hua Zhou, Jin Zhou, and Wei Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13, 2017.

[145] Paulo Guimaraes, Richard Lindrooth, et al. Dirichlet-multinomial regression. *Economics Working Paper Archive at WUSTL, Econometrics*, 509001, 2005.

[146] XuGang Wang, Hongan Wang, Guozhong Dai, and Zheng Tang. A reliable resilient back-propagation method with gradient ascent. In De-Shuang Huang, Kang Li, and George W. Irwin, editors, *Computational Intelligence, International Conference on Intelligent Computing, ICIC 2006, Kunming, China, August 16-19, 2006. Proceedings, Part II*, volume 4114 of *Lecture Notes in Computer Science*, pages 236–244. Springer, 2006.

[147] Xinyi Wang and Yi Yang. Neural topic model with attention for supervised learning. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 1147–1156. PMLR, 2020.

[148] Tzu-Tsung Wong. Alternative prior assumptions for improving the performance of naïve bayesian classifiers. *Data Mining and Knowledge Discovery*, 18(2):183–213, 2009.

[149] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[150] Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of bionlp shared task 2013. In Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum, editors, *Proceedings of the BioNLP Shared Task 2013 Workshop, Sofia, Bulgaria, August 9, 2013*, pages 1–7. Association for Computational Linguistics, 2013.

[151] David C Blair and Melvin E Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

[152] Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157, 2018.

[153] S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pages 267–281, 2021.

[154] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[155] Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 2023.

[156] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[157] Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498, 2022.

[158] Zhiyuan Chen and Bing Liu. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1116–1125, 2014.

[159] Fatma Najar and Nizar Bouguila. Exact fisher information of generalized dirichlet multinomial distribution for count data modeling. *Inf. Sci.*, 586:688–703, 2022.

[160] 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, 2011.

[161] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*, volume 48. IEEE Transactions, 2001.

[162] Nizar Bouguila. Deriving kernels from generalized dirichlet mixture models and applications. *Information processing & management*, 49(1):123–137, 2013.

[163] Ali Shojaee Bakhtiari and Nizar Bouguila. A novel hierarchical statistical model for count data modeling and its application in image classification. In *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part II*, volume 7664 of *Lecture Notes in Computer Science*, pages 332–340. Springer, 2012.

[164] Nuha Zamzami and Nizar Bouguila. High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution. *Inf. Sci.*, 524:116–135, 2020.

[165] Rameshwar D Gupta and Debasis Kundu. Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(1):117–130, 2001.

[166] W Buntine. Computation with the exponential family and graphical models. *unpublished handouts, NATO Summer School on Graphical Models, Erice, Italy*, 25, 1996.