*electronics*

MDPI

*Article*

# Hidden Variable Models in Text Classification and Sentiment Analysis

**Pantea Koochemeshkian *, Eddy Ihou Koffi and Nizar Bouguila**

Concordia Institute for Information Systems Engineering (CIISE), Montreal, QC H3G 1M8, Canada; k_ihou@encs.concordia.ca (E.I.K.); nizar.bouguila@concordia.ca (N.B.)
* Correspondence: p_kooche@encs.concordia.ca

**Abstract:** In this paper, we are proposing extensions to the multinomial principal component analysis (MPCA) framework, which is a Dirichlet (Dir)-based model widely used in text document analysis. The MPCA is a discrete analogue to the standard PCA (it operates on continuous data using Gaussian distributions). With the extensive use of count data in modeling nowadays, the current limitations of the Dir prior (independent assumption within its components and very restricted covariance structure) tend to prevent efficient processing. As a result, we are proposing some alternatives with flexible priors such as generalized Dirichlet (GD) and Beta-Liouville (BL), leading to GDMPCA and BLMPCA models, respectively. Besides using these priors as they generalize the Dir, importantly, we also implement a deterministic method that uses variational Bayesian inference for the fast convergence of the proposed algorithms. Additionally, we use collapsed Gibbs sampling to estimate the model parameters, providing a computationally efficient method for inference. These two variational models offer higher flexibility while assigning each observation to a distinct cluster. We create several multitopic models and evaluate their strengths and weaknesses using real-world applications such as text classification and sentiment analysis.

**Keywords:** multinomial PCA; generalized Dirichlet MPCA; Beta-Liouville MPCA; topic modeling; text classification; sentiment analysis; variational inference; collapsed Gibbs sampling; dimensionality reduction; text clustering



**Citation:** Koochemeshkian, P.; Ihou Koffi, E.; Bouguila, N. Hidden Variable Models in Text Classification and Sentiment Analysis. *Electronics* **2024**, *13*, 1859. https://doi.org/10.3390/electronics13101859

Academic Editors: Ruifeng Xu and Praveen Kumar Donta

Received: 25 March 2024
Revised: 30 April 2024
Accepted: 7 May 2024
Published: 10 May 2024



## 1. Introduction

In this fast-paced world of technological advances, one of the most significant contributing factors has been the emergence of various digital data forms, opening opportunities in different fields to gather helpful information. Everyday, massive amounts of digital data are stored in digital data archives. The same distinction can be made for the enormous quantity of textual data available on the Internet. Therefore, it is critical to developing effective and scalable statistical models to extract hidden knowledge from such rich data [1].

One of the main challenges in the statistical analysis of textual data is capturing and representing their complexity. Different approaches have been applied to deal with this problem. Furthermore, due to information technology's rapid development, vast quantities of scientific documents are now freely available to be mined. Thus, the analysis and mining of scientific documents have been very active research areas for many years.

Data projection and clustering are crucial for document analysis, with projections aimed at creating low-dimensional, meaningful data representations and clustering and grouping similar data patterns [2,3]. Traditionally, these methods are studied separately, but they intersect in many applications [3]. 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 [4], summarization [5], segmentation [6], topic modeling [7], and sentiment analysis [8].

In this paper, we will focus 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 [9]. The matrix decomposition approach, such as probabilistic latent semantic analysis (PLSA) [10,11], analyzes text via mining and requires a deep understanding of the corpus structure. PLSA, also known as probabilistic latent semantic indexing (pLSI) [11], 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, and a document is treated as a combination of various topics, each represented by a random vector of words [3].

Meanwhile, the research by [12] 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) [12] was introduced, enhancing pLSI using a Dirichlet distribution for topic mixtures. LDA stands out as a more effective generative model, though it still lacks robust clustering capabilities [3]. 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 [13,14].

The LDA model [12] has been proposed to solve these shortcomings, and this model has proved to be an efficient and scalable data processing method [15,16].

The main issue with current text analysis models is their failure to clearly define a probability model encompassing hidden variables and assumptions [11,17–19]. 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 [19,20]. Extensions of LDA, like its hierarchical [21] and online versions [22], 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) methods [23], which, despite their effectiveness, require longer convergence times compared to the variational Bayes approach.

In this paper, 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 [16,24,25]. Both models employ variational Bayesian inference and collapsed Gibbs sampling for efficient and scalable computational performances, which is critical for handling large datasets.

The generalized Dirichlet (GD) distribution, introduced in [26], 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 was validated through a rigorous empirical evaluation on real-world datasets, which demonstrated 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.

The structure of the rest of this paper is as follows. In Section 2, we cover the related work. Section 3 introduces the extension of MPCA with generalized Dirichlet and Beta-Liouville distributions with all the details about the parameters estimation. Section 4 is devoted to the discussion of the experimental results. Finally, we conclude our work in Section 6.

## 2. Related Work

In this section, we delve into the vast array of the literature on topic modeling approaches. The foundation of this field is built upon traditional topic modeling tech-

niques [10,11], with significant contributions from topic-class modeling [27–29] and the nuanced exploration of global and local document features [30,31].

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 [32]. This approach provides a nuanced lens with which to view the thematic undercurrents of scholarly communication.

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 [33]. 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 [34]. These studies, which utilize convolutional neural networks and Inception-V3 transfer learning [35], pave the way for multimodal sentiment analysis, potentially influencing strategies for textual sentiment analysis.

A hybrid deep learning method has been introduced for analyzing sentiment polarities and knowledge graph representations, particularly focusing on health-related social media data, like tweets on monkeypox [36]. 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 [37], continue to be pivotal. Probabilistic latent semantic indexing (pLSI) [11] and latent Dirichlet allocation (LDA) [12] further enrich the discussion on discrete data and topic modeling. Non-negative matrix factorization (NMF) [17] has also demonstrated effectiveness, emphasizing the need for models that can simultaneously handle clustering and projection. In 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 [19].

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, reflecting the ongoing dialogue within the research community on these topics.

### *2.1. Multinomial PCA*

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

In the probabilistic PCA (pPCA) framework, as detailed in the work in [38], 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:

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

where $Z$ is a "standard" normal distribution, ($\beta$, $\mu$) are the model parameters, and $\sigma^2$ is the variance that is learned using maximum likelihood inference.

The Gaussian assumption is suitable for real-valued data, yet it is less applicable to non-negative count data. Addressing this, [19] 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 $\alpha = (\alpha_1, \ldots, \alpha_k) \geq 0$.

Then, the probabilistic function is assumed to be multinomial:

$$\begin{aligned} m &\sim Dirichlet(\alpha) \\ C &\sim Multinomial(m, L), \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{3}$$

The variables $m$ and $w$ are assumed to be hidden parameters for each document. For the parameter estimation of MPCA, first, the variable $\Omega$ is estimated using the Dirichlet prior on $m$ using parameters $\alpha$ [19]. The likelihood model for the MPCA is given as follows [20]:

$$p(m, w|\alpha, \Omega) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} C^L_{w_{1,1}, w_{1,2} \ldots, w_{k,1}, w_{1,J} \ldots w_{K,J}} \prod_K m_K^{a_k - 1} \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \tag{4}$$

In the MPCA model, it is assumed that each observation $x_i$ can be broken down into a probabilistic mixture of $K$ topics that represent the whole corpus. Then, $m$ indicates the observation with mixture weights in the latent space, and $\Omega$ is a global parameter that encapsulates all the information at the corpus level.

As a result, the following equation is derived when the hidden variables have a Dirichlet prior [19]:

$$\begin{aligned} m &\sim Dirichlet(\alpha) \\ \Omega_k &\sim Dirichlet(2f) \end{aligned} \tag{5}$$

The following updated formula converges to the local maximum $\log p(\Omega, \alpha_m|r)$, where $\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}$ 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 [19]:

$$\gamma_{j,k,[i]} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{1}{\Omega_{k,j} m_{k,[i]}} \tag{6}$$

$$m_{k,[i]} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \left(a_k - 1 + \sum_j r_{j,[i]} \gamma_{j,k,[i]}\right) \tag{7}$$

Equations (8) and (9) are the parameters for a multinomial and a Dirichlet, respectively.

$$\Omega_{k,j} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \left(2f + \sum_i r_{j,[i]} \gamma_{j,k,[i]}\right) \tag{8}$$

$$\Psi_0(a_k) - \Psi_0\left(\sum_k a_k\right) = \frac{\log(1/k) + \sum_i \log(m_{k,[i]})}{1 + I} \tag{9}$$

According to the exponential family definition (Appendix A), Equation (9) rewrites $\alpha$ in terms of its dual representation. Minka's approach is used to derive $\alpha$, where $n_k$ is the number of times that the outcome was $k$ [39]:

$$\begin{aligned} n_k &= \sum_i \delta(x_i - k) \\ n_i &= \sum_k n_i k \end{aligned} \tag{10}$$

$$\alpha_k^{new} = a_k \frac{\sum_i \Psi(n_{ik} + a_k) - \Psi(a_k)}{\sum_i \Psi(n_i k + \sum_k a_k) - \Psi(\sum_k a_k)} \tag{11}$$

Connection between MPCA and LDA

The multinomial PCA model is closely connected to LDA [12] 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 $w_i = w_{in}$, where $n = 1 \ldots L_i$. 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:

$$\begin{aligned} &\theta_k \sim Dirichlet(\alpha) \\ &z_{in}|\theta_i \sim Multinomial(1, \theta_i) \\ &w_{in}|z_{ink} \sim Multinomial(1, \beta_k) \end{aligned} \tag{12}$$

At the word level, marginalizing on $z_{in}$ yields a distribution similar to Equation (3):

$$w_{in}|\theta_i \sim Multinomial(1, \beta\theta_i) \tag{13}$$

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

## 3. Proposed Models

In this section, we present two pioneering models, generalized Dirichlet Multinomial Principal Component Analysis (GDMPCA) and Beta-Liouville Multinomial Principal Component Analysis (BLMPCA), which were designed to revolutionize text classification and sentiment analysis. At the core of our approaches is the integration of generalized Dirichlet and Beta-Liouville 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 performances of GDMPCA and BLMPCA in terms of accuracy, adaptability, and computational efficiency, as evidenced by an 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 1 summarizes the relevant variables for the proposed models.

**Table 1.** Parameters of generalized Dirichlet and Beta-Liouville distributions.

| Parameter | Generalized Dirichlet (GDMPCA) | Beta-Liouville (BLMPCA) |
|---|---|---|
| $\xi$ | Parameters of GD distribution | Not applicable |
| Y | 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 |
| $\Omega$ | Multinomial parameters (words) | Multinomial parameters (words) |
| $L$ | Number of words per document | Number of words per document |
| $C, \Omega_k, c_k$ | Multinomial parameters for topics | Multinomial parameters for topics |

### *3.1. Generalized Dirichlet Multinomial PCA*

Bouguila [40] demonstrated that when mixture models are used, the generalized Dirichlet (GD) distribution is a reasonable alternative to the Dirichlet distribution for clustering count data.

As we mentioned previously, the GD distribution, like the Dirichlet distribution, is a conjugate prior to the multinomial distribution. Furthermore, the GD has a more general covariance matrix [40].

Therefore, the variational Bayes approach will be utilized to develop an extension of the MPCA model incorporating the generalized Dirichlet assumption. GDMPCA is anticipated to perform effectively because the Dirichlet distribution is a specific instance of the GD [41]. Like MPCA, GDMPCA is a fully generative model applied to a corpus. It considers a collection of $M$ documents represented as the corpus, denoted by $D = \{w_1, w_2, \ldots, w_M\}$. Each document $w_m$ consists of a sequence of $N_m$ words, expressed as $w_m = (w_{m1}, \ldots, w_{mN_m})$. Words within a document are represented by binary vectors from a vocabulary of $V$ words, where if the $j$-th word is selected, $w_j^n = 1$, and if not, $w_j^n = 0$ [42]. The GDMPCA model then describes the generation of each word in the document through a series of steps involving $c$, a $d+1$ dimensional binary vector of topics:

$$\begin{aligned} m &\sim GD(\xi) \\ z &\sim Multinomial(m, L) \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{14}$$

If the i-th topic is chosen, $z_i^n = 1$, and 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 $\gamma_i = b_i - a_{i+1} - b_{i+1}$ [42]:

$$p(m_1, \ldots, m_d|\xi) = \prod_{i=1}^{d} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} m_i^{a_l - 1} (1 - \sum_{j=1}^{i} m_j)^{\gamma_i} \tag{15}$$

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

The mean and the variance matrix of the GD distribution are as follows [41]:

$$E(m_i) = \frac{a_l}{a_l + b_l} \prod_{k=1}^{i-1} \frac{b_k}{a_k + b_k} \tag{16}$$

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

and the covariance between $m_i$ and $m_j$ is given by

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

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 [43]. The Dirichlet and GD distributions are both members of the exponential family (Appendix A). 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:

$$\begin{aligned} p(m, w|\xi, \Omega) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} z^L_{w_{1,1}, w_{1,2} \ldots, w_{k,1}, w_{1,J} \ldots w_{K,J}} m_k^{b_{k-1}-1} \prod_{i=1}^{k-1} [m_i^{a_{i-1}} \\ (\sum_{j=1}^{k} m_j)^{b_{i-1}+(a_i+b_i)}] \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \end{aligned} \tag{19}$$

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:

$$\begin{aligned} m &\sim GD(\xi) \\ \Omega_k &\sim GD(2f) \end{aligned} \tag{20}$$

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

First, we will calculate the parameters of GD utilizing the Hessian matrix as described in Appendix B.1.2, following Equations (19) and (20). To find the optimal variational parameters, we minimize the Kullback–Leibler (KL) divergence between the variational distribution and the posterior distributions $p(m, w|\Omega, \xi)$. This is achieved through a repetitive fixed-point method. We specify the variational parameters as follows:

$$q(m, c|\gamma, \Phi) = q(m|\gamma)\prod_{k=1}^{K} q(c_k|\Phi_k) \tag{21}$$

As an alternative to the posterior distribution $p(m, c, w, \xi, \Omega)$, we determine the variational parameters $\gamma$ and $\Phi$ 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 parameters $\gamma$ and $\Phi$ [44]:

$$\begin{aligned} \log p(w|\xi, \Omega) &= \log \int \sum_z p(m, c, w|\xi, \Omega) dm \\ &= \log \int \sum_z \frac{p(m, c, w|\xi, \Omega) q(m, c)}{q(m, c)} dm \\ &\geq \int \sum_z \log p(m, c, w|\xi, \Omega) q(m, c) dm \\ &- \int \sum_z q(m, c) \log q(m, c) dm \\ &= E[\log p(m, c, w|\xi, \Omega)] - E[\log q(m, c)] \end{aligned} \tag{22}$$

Consequently, Jensen's inequality provides a lower bound on the log likelihood for any given variational distribution $q(m, c|\gamma, \Phi)$.

If the right-hand side of Equation (22) is denoted as $\mathcal{L}(\gamma, \Phi; \xi, \Omega)$, the discrepancy between the left and right sides of this equation represents the KL divergence between the variational distribution and the true posterior probabilities. This re-establishes the importance of the variational parameters, leading to the following expression:

$$\log p(w|\xi, \Omega) = \mathcal{L}(\gamma, \Phi; \xi, \Omega) + D(q(m, c|\gamma, \Phi))|p(m, c|x, \xi, \Omega) \tag{23}$$

As demonstrated in Equation (23), maximizing the lower bound $\mathcal{L}(\gamma, \Phi; \xi, \Omega)$ with respect to $\gamma$ and $\Phi$ is equivalent to minimizing the Kullback–Leibler (KL) divergence between the variational posterior probability. By factorizing the variational distributions, we can describe the lower bound as follows:

$$\begin{aligned}\mathcal{L}(\gamma, \Phi; \xi, \Omega) = E_q[\log p(m|\xi)] + E_q[\log p(c|m)] + E_q[\log p(w|c, \Omega)] \\ - E_q[\log q(m)] - E_q[\log q(c)]\end{aligned} \tag{24}$$

After that, we can extend Equation (A7) in terms of the model parameters $(\xi, \Omega)$ and variational parameters $(\gamma, \Phi)$ (A13).

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

$$\begin{aligned}\mathcal{L}[m_{nl}] = m_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi)) + m_{nl} \log \Omega_{w(lv)} - m_{nl} \log m_{nl} \\ + \lambda_n(\sum_{ll=1}^{d+1} m_{n(ll)} - 1)\end{aligned} \tag{25}$$

and therefore, we have

$$\frac{\partial \mathcal{L}}{\partial \phi_{nl}} = (\Psi(\gamma_l) - \Psi(\gamma_l + \Phi)) + \log \Omega_{lv} - \log \phi_{nl} - 1 + \lambda_n \tag{26}$$

Setting the above equation to zero leads to

$$m_{nl} = \Omega_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi))} \tag{27}$$

Next, we maximize Equation (A13) with respect to $\gamma_i$. The terms containing $\gamma_i$ are

$$\begin{aligned}\mathcal{L}[\xi_q] = \sum_{l=1}^{d}[a_l(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi)) + (\Psi(\gamma_l) - \\ \Psi(\gamma_l + \Phi))(b_l - a_{l+1} - b_{l+1})] \\ + \sum_{n=1}^{N} m_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi) + \sum_{n=1}^{N} m_{n(d+1)}(\Psi(\gamma_d) - \Psi(\gamma_d + \Phi_d)) \\ - \sum_{l=1}^{d}(\log \Gamma(\gamma_l + \Phi_l) - \log \Gamma(\gamma_l) - \log \Gamma(\Phi_l)) \\ + \sum_{l=1}^{d}(\Psi(\gamma_l) - \gamma_l(\Psi(\gamma_l + \Phi_l)) \\ + (\Psi(\Phi) - \Psi(\Phi + \gamma_l))(\Phi - \gamma_{l+1} - \Phi_{l+1})))\end{aligned} \tag{28}$$

Setting the derivative of the above equation to zero leads to the following updated parameters:

$$\gamma_l = a_l + \sum_{n=1}^{N} m_{nl} \tag{29}$$

$$\Phi_l = b_l + \sum_{n=1}^{N} \sum_{l=l+1}^{d+1} m_{n(l)} \tag{30}$$

The challenge of deriving empirical Bayes estimates for the model parameters $\xi$ and $\Omega$ is tackled by utilizing the variational lower bound as a substitute for the marginal log probability, using variational parameters $\gamma$ and $\Phi$. The empirical Bayes estimates are then determined by maximizing this lower bound in relation to the model parameters. Until now, our discussion has centered on the log probability for a single document; the overall variational lower bound is computed as the sum of the individual lower bounds from each document. In the M-step, this bound is maximized with respect to model parameters $\xi$ and $\Omega$. Consequently, the entire process is akin to performing a coordinate ascent as outlined in Equation (31). We formulate the update equation for estimating $\Omega$ by isolating terms and incorporating Lagrange multipliers to maximize the bound with respect to $\Omega$:

$$\mathcal{L}[\Omega] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{K+1} \sum_{j=1}^{V} m_{dnl} w_{dn}^{j} \log \Omega_{(lj)} + \sum_{l=1}^{K+1} \lambda_l \Big( \sum_{j=1}^{V} \Omega_{w(ij)} \Big) \tag{31}$$

To derive the update equation for $\Omega_{(lj)}$, we take the derivative of the variational lower bound with respect to $\Omega_{(lj)}$ and set this derivative to zero. This step ensures that we find the point where the lower bound is maximized with respect to the parameter $\Omega_{(lj)}$.

$$\Omega_{(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j} \tag{32}$$

The updates mentioned lead to convergence at a local maximum of the lower bound of $\log p(\Omega, \xi | r)$, which is optimal 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.

$$\Phi_l = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} m_{nl} (\Psi(\gamma_l) - \Psi(\gamma_l + \Phi)) \tag{33}$$

$$\gamma_l = a_l + \sum_{n=1}^{N} m_{nl} \tag{34}$$

$$\Omega_{(lj)} = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} (2 f_j \sum_{d=1}^{M} \sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j}) \tag{35}$$

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:

$$p(c, z, \theta, \varphi, w|, \Omega, \xi, \mu) = p(w|\mu) p(\theta|\Omega) p(\varphi|\xi) \times \prod_{n=1}^{N} p(z_n|\theta) p(x_n|z_{nn}, \varphi) \tag{36}$$

Here, $p(\theta|\Omega)$ signifies the GD document prior distribution, where $\Omega = (a_1, b_1, \ldots, a_n, b_n)$ serves as a 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 follows:

$$p(c, z|w, \Omega, \xi) = W \int_{\theta} \int_{\varphi} p(c, z, \theta, \varphi, |\Omega, \xi) d\varphi d\theta \tag{37}$$

Crucially, the joint distribution is expressed as a product of Gamma functions, as highlighted in prior research [12,45,46]. This expression facilitates the determination of the expectation value for the accurate posterior distribution:

$$p(z_{ij} = k|c, w, \Omega, \xi) = \mathrm{E}_{p(z^{-ij}|w,c,\Omega,\xi)}[p(z_{ij} = k|z^{-ij}, c, w, \Omega, \xi)] \tag{38}$$

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

$$\begin{aligned} p(z_{ij} = k|z^{-ij}, c, w, \Omega, \xi) \propto & \left[\frac{(N_{jk}^{-ij} + \alpha_{wk})(\beta_{wk} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})}{(\alpha_{wk}\beta_{wk} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})}\right] \\ & \times \left[\frac{(N_{kv_{ij}}^{-ij} + a_v)(b_v + \sum_{d=v}^{V+1} N_{kd_{ij}}^{-ij})}{a_v + b_v + \sum_{d=v}^{V+1} N_{kd_{ij}}^{-ij})}\right] = A(K) \end{aligned} \tag{39}$$

This leads to a posterior probability normalization as follows:

$$p(z_{ij} = k|z^{-ij}, x, \Omega, \xi) = \frac{A(k)}{\sum_{k'=1}^{K} A(k')} \tag{40}$$

The sequence from Equation (38) to Equation (40) delineates the complete collapsed Gibbs sampling procedure, encapsulated as follows:

$$p(z_{ij} = k|c, w, \Omega, \xi) = \mathrm{E}_{p(z^{-ij}|w,c,\Omega,\xi)}\left[\frac{A(k)}{\sum_{k'=1}^{K} A(k')}\right] \tag{41}$$

The implementation of collapsed Gibbs sampling in our GD-centric model facilitates sampling directly from the actual posterior distribution $p$, as indicated in Equation (41). This sampling technique is deemed more accurate than those employed in variational inference models, which typically approximate the distribution from which samples are drawn [46,47]. Hence, our model's precision is ostensibly superior.

Upon the completion of the sampling phase, parameter estimation is conducted using the methodologies discussed.

### 3.2. Beta-Liouville Multinomial PCA

For the 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 for every single word of the document. The BLMPCA model proceeds with generating every single word of the document with the following steps, where $c$ is a $d+1$-dimensional binary vector of topics defined:

$$\begin{aligned} m &\sim BL(\Upsilon) \\ z &\sim Multinomial(m, L), \\ w_k &\sim Multinomial(\Omega_k, c_k) \end{aligned} \tag{42}$$

In the model described, each topic is represented by a binary variable, where $z_i^n = 1$ indicates that the i-th topic is chosen for the n-th word, and $z_i^n = 0$ indicates it is not chosen. The vector $z_n$ is a $(D+1)$-dimensional binary vector representing the topic assignments across all $D+1$ topics for a given word. The vector $m$ is defined as $m = (m_1, m_2, \ldots, m_{D+1})$, where $m_{D+1} = 1 - \sum_{i=1}^{D} m_i$ captures the distribution of topic proportions across the document, ensuring that the sum of proportions across all topics equals 1.

A chosen topic is associated with a multinomial prior $w$ over the vocabulary, where $\Omega_{w_{ij}} = p(w^j = 1|z^i = 1)$ describes the probability of the j-th word being selected given that the i-th topic is chosen. This formulation allows for each word in the document to be drawn randomly from the vocabulary conditioned on the assigned topic.

The probability $p(w_n|z_n, \Omega_w)$ is a multinomial probability that conditions on the topic assignments $z_n$ and the topic-word distributions $\Omega_w$, effectively modeling the likelihood of each word in the document given the topic assignments.

Additionally, $BL(\Upsilon)$ represents a d-variate Beta-Liouville distribution with parameters $\Upsilon = (\alpha_1, ..., \alpha_D, \alpha, \beta)$. 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.

$$P(\theta_1, \ldots, \theta_D|\Upsilon) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{\theta_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times \left(\sum_{d=1}^{D} \theta_d\right)^{\alpha - \sum_{l=1}^{D} \alpha_l} \times \left(1 - \sum_{l=1}^{D} \theta_l\right)^{\beta - 1} \tag{43}$$

A Dirichlet distribution is the special case of BL if $\beta_d = \alpha_{d+1} + \beta_{d+1}$ [42,45]. The mean, the variance, and the covariance in the case of a BL distribution are as follows [45]:

$$E(\theta_d) = \frac{\alpha}{\alpha+\beta} \frac{\alpha_d}{\sum_{d=1}^{D} \alpha_d} \tag{44}$$

$$var(\theta_d) = \left(\frac{\alpha}{\alpha+\beta}\right)^2 \frac{\alpha_d(\alpha_d+1)}{(\sum_{m=1}^{D} \alpha_m)(\sum_{m=1}^{D} \alpha_m + 1)} - E(\theta_d)^2 \frac{\alpha_d^2}{(\sum_{m=1}^{D} \alpha_m)^2} \tag{45}$$

and the covariance between $\theta_l$ and $\theta_k$ is given by

$$Cov(\theta_l, \theta_k) = \frac{\alpha_l \alpha_k}{\sum_{d=1}^{D} \alpha_d} \left( \frac{\frac{(\alpha+1)(\alpha)}{(\alpha+\beta+1)(\alpha+\beta)}}{\sum_{d=1}^{D} \alpha_d + 1} - \frac{\frac{\alpha}{\alpha+\beta}}{\sum_{d=1}^{D} \alpha_d} \right) \tag{46}$$

The earlier equation illustrates that the covariance matrix of the Beta-Liouville distribution offers a broader scope compared to the covariance matrix of the Dirichlet distribution. For the parameter estimation of BLMPCA, first, the parameter $\Omega$ is estimated using the Beta-Liouville prior on $m$ using parameter $\Upsilon$ [19]. The likelihood model for the BLMPCA is given as follows:

$$p(m, w|\Upsilon, \Omega) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)} z^{L}_{w_{1,1}, w_{1,2} \ldots, w_{k,1}, w_{1,J} \ldots w_{K,J}} \Big[ \frac{1}{\Gamma(\alpha_d)} m_k^{\alpha_d - 1} + \sum_k m_k^{\alpha - \sum_d \alpha_d} + (1 - \sum_k m_k)^{\beta - 1} \Big] \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \tag{47}$$

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

$$\begin{aligned} m &\sim BL(\Upsilon) \\ \Omega_k &\sim BL(2f) \end{aligned} \tag{48}$$

In the following step, we will estimate the parameters for $\Omega$ using the Beta-Liouville prior and the Hessian matrix (Appendix C). As we explained in the previous Section 3.1, we should estimate the model parameters $(\Upsilon, \Omega)$ and the variational parameters $(\gamma, \Phi)$ according to Equations (21), (22) and (A7) to find $m_{nl}$, and we proceed to maximize with respect to $m_{nl}$; so, we have the following equations:

$$\begin{aligned}
\mathcal{L}(\gamma, \Phi; \Upsilon, \Omega) = & \log(\Gamma(\sum_{d=1}^{D} \alpha_d)) + \log(\Gamma(\alpha + \beta)) - \log(\Gamma(\alpha)) \\
& - \log(\Gamma(\beta)) - \sum_{d=1}^{D} \log \Gamma(\alpha_d) + \sum_{d=1}^{D} \alpha_d (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l) \\
& + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(\Psi(\beta_\gamma) \\
& - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N} \sum_{d=1}^{D} m_{nd} (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N} m_{n(D+1)} (\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\
& + \sum_{n=1}^{N} \sum_{l=1}^{D+1} \sum_{j=1}^{V} m_{nl} w_n^j \log(\Omega_{lj}) \\
& - \big( \log(\Gamma(\sum_{l=1}^{D} \alpha_l)) + \log(\Gamma(\alpha + \beta)) - \log \Gamma(\alpha) - \log \Gamma(\beta) \\
& - \sum_{i=1}^{D} \log \Gamma(\alpha_i) \\
& + \sum_{i=1}^{D} \alpha_i (\Psi(\gamma_{mi}) - \Psi(\sum_{l=1}^{D} \gamma_{m(l)})) + \alpha_( \Psi(\alpha_{m\gamma}) \\
& - \Psi(\alpha_{m\gamma} \beta_{m\gamma})) + \beta(\Psi(\beta_{m\gamma}) - \Psi(\alpha_{m\gamma} + \beta_{m\gamma}))\big) \\
& - \big( \sum_{n=1}^{N} \sum_{l=1}^{D+1} m_{nl} \log(m_{nl}) \big)
\end{aligned} \tag{49}$$

To find $m_{nl}$ , we proceed to maximize with respect to $\phi_{nl}$:

$$\begin{aligned}
\mathcal{L}[m_{nl}] = & m_{nl}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D} \gamma_l)) + m_{nl} \log \beta_{w(iv)} - m_{nl} \log(m_{nl}) \\
& + \lambda_n (\sum_{l=1}^{D} m_{nl} - 1)
\end{aligned} \tag{50}$$

Therefore, we have

$$\frac{\partial \mathcal{L}}{\partial \phi_{nl}} = (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \log \beta_{w(iv)} - \log \phi_{nl} - 1 + \lambda_n \tag{51}$$

The next step is to optimize Equation (49) to find the update equations for the variational; we separate the terms containing the variational Beta-Liouville parameters once more.

$$\begin{aligned}\mathcal{L}[\xi_q] = {} & \alpha_d(\Psi(\gamma_d)) - \Psi(\sum_{l=1}^{D}\gamma_l) + \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma \\ & + \beta_\gamma)) + \beta(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ & + \sum_{n=1}^{N}\phi_n(\Psi(\gamma_l) - \Psi(\sum_{l=1}^{D}\gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ & + \sum_{n=1}^{N}\phi_{n(D+1)}(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ & - (\log(\Gamma(\sum_{l=1}^{D}\gamma_l)) + \log(\gamma(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) \\ & - \log(\Gamma(\beta_\gamma)) - \log(\Gamma(\gamma_l)) \\ & + \gamma_l(\Psi(\gamma_l) + \Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) - \Psi(\sum_{l=1}^{D}\gamma_l) \\ & + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ & + \beta_\gamma(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)))\end{aligned} \tag{52}$$

Selecting the terms containing variational Beta-Liouville variables $\gamma_i, \alpha_\gamma$, and $\beta_\gamma$, we have

$$\begin{aligned}\mathcal{L}(\gamma_i) = {} & \alpha_i(\Psi(\gamma_i)) - (\sum_{l=1}^{D}\alpha_l)(\Psi(\sum_{l=1}^{D}\gamma_l)) + \sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D}\gamma_l)) \\ & - (\log\Gamma(\sum_{l=1}^{D}) - \log\Gamma(\gamma_i) + \gamma_i(\Psi(\sum_{l=1}^{D}\gamma_l)\sum_{d=1}^{D}\gamma_d)\end{aligned} \tag{53}$$

and

$$\begin{aligned}\mathcal{L}[\alpha_\gamma] = {} & \alpha(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) + \beta(-\Psi(\alpha_\gamma + \beta_\gamma)) \\ & + (\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))\sum_{n=1}^{N}\sum_{i=1}^{D}\phi_{ni}\sum_{n=1}^{N}\phi_{n(D+1)}(-\Psi(\alpha_\gamma + \beta_\gamma)) \\ & - (\log(\alpha_\gamma + \beta_\gamma) - \log(\Gamma(\alpha_\gamma)) + \alpha_\gamma(\Psi(\alpha_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \\ & + \beta_\gamma(-\Psi(\alpha_\gamma + \beta_\gamma)))\end{aligned} \tag{54}$$

Setting Equations (52)–(54) to zero, we have the following update parameters:

$$\gamma_i = \alpha + \sum_{n=1}^{N}\phi_{ni} \tag{55}$$

$$\alpha_\gamma = \alpha + \sum_{n=1}^{N}\sum_{d=1}^{D}\phi_{nd} \tag{56}$$

$$\beta_\gamma = \beta + \sum_{n=1}^{N}\phi_{n(D+1)} \tag{57}$$

We address the challenge of deriving empirical Bayes estimates for the model parameters $\Upsilon$ and $\Omega$ by utilizing the variational lower bound as a substitute for the marginal log likelihood. This approach fixes the variational parameters $\gamma$ and $\Phi$ at values determined through variational inference. We then optimize this lower bound to obtain the empirical Bayes estimates of the model parameters.

To estimate $\Omega_w$, we formulate necessary update equations. The process of maximizing Equation (52) with respect to $\Omega$ results in the following equation:

$$\mathcal{L}[\Omega_w] = \sum_{d=1}^{M}\sum_{n=1}^{N_s}\sum_{l=1}^{D+1}\sum_{j=1}^{V}\phi_{dnl}w_{dn}^{j}\log(\Omega_{w(lj)}) + \sum_{l=1}^{D+1}\lambda_l(\sum_{j=1}^{V}\Omega_{w(lj)} - 1) \tag{58}$$

Taking the derivatives with the respect to $\beta_{w(lj)}$ and setting it to zero yields in Appendix C.1:

$$\Omega_{w(lj)} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} m_{dnl}w_{dn}^{j} \tag{59}$$

Beta-Liouville Parameter

The objective of this subsection is to determine the estimates of the model's parameters using variational inference techniques [48].

$$\begin{aligned}\mathcal{L}[\xi] = &\sum_{m=1}^{M}(\log(\Gamma(\sum_{l=1}^{D}\alpha_l)) + \log(\Gamma(\alpha+\beta)) - \log\Gamma(\alpha) - \log\Gamma(\beta)\\ &- \sum_{i=1}^{D}\log\Gamma(\alpha_i) + \sum_{i=1}^{D}\alpha_i(\Psi(\gamma_{mi}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)}))\\ &+ \alpha(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}\beta_{m\gamma})) + \beta(\Psi(\beta_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma})))\end{aligned} \tag{60}$$

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

$$\begin{aligned}\frac{\partial\mathcal{L}[\xi]}{\partial\alpha_l} &= M(\Psi(\sum_{l=1}^{D}) - \Psi(\alpha_l)) + \sum_{m=1}^{M}(\Psi^{'}(\gamma_{ml}) - \Psi(\sum_{l=1}^{D}\gamma_{m(l)}))\\ \frac{\partial\mathcal{L}[\xi]}{\partial\alpha} &= M[\Psi(\alpha+\beta) - \Psi(\alpha)] + \sum_{m=1}^{M}(\Psi(\alpha_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma}))\\ \frac{\partial\mathcal{L}[\xi]}{\partial\beta} &= M[\Psi(\alpha+\beta) - \Psi(\beta)] + \sum_{m=1}^{M}(\Psi(\beta_{m\gamma}) - \Psi(\alpha_{m\gamma}+\beta_{m\gamma}))\end{aligned} \tag{61}$$

From the equations presented earlier, it is evident that the derivative in Equation (52) 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 [49]:

$$\begin{aligned}\frac{\partial^2\mathcal{L}[\xi]}{\partial\alpha_l\alpha_j} &= M(-\delta(i,j)\Psi^{'}(\alpha_i) + \Psi^{'}(\sum_{l=1}^{D}\alpha_l))\\ \frac{\partial^2\mathcal{L}[\xi]}{\partial\alpha^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\alpha))\\ \frac{\partial^2\mathcal{L}[\xi]}{\partial\alpha\partial\beta} &= M\Psi^{'}(\alpha+\beta)\\ \frac{\partial^2\mathcal{L}[\xi]}{\partial\beta^2} &= M(\Psi^{'}(\alpha+\beta) - \Psi^{'}(\beta))\end{aligned} \tag{62}$$

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 using parameters $\alpha_d$, $\alpha$, and $\beta$. Each of the two parts' parameter derivation will be identical to the Newton–Raphson model provided by MPCA and GDMPCA.

### 3.3. Inference via Collapsed Gibbs Sampling

The collapsed Gibbs sampler (CGS) contributes to the inference by estimating posterior distributions through a Bayesian network of conditional probabilities, which are determined through a sampling process of hidden variables. Compared to the traditional Gibbs sampler that functions in the combined space of latent variables and model parameters, the CGS offers significantly faster estimation times. The CGS operates within the collapsed space of latent variables, where, in the joint distribution $p(X, z, \theta, \phi, w|\Omega, \Upsilon, \mu)$, the model parameters $\theta$ and $\phi$ are marginalized out. This marginalization leads to the marginal joint distribution $p(X, z, w|\Omega, \Upsilon, \mu)$, which is defined as follows:

$$p(x, z, w|\Omega, \Upsilon) = W \int_{\theta} \int_{\varphi} p(X, z, \theta, \varphi, w|\Omega, \xi) d\varphi d\theta \tag{63}$$

In using Equation (63), the method calculates the conditional probabilities of the latent variables $z_{ij}$ by considering the current state of all other variables while excluding the specific variable $z_{ij}$ itself [50]. Meanwhile, the collapsed Gibbs sampler (CGS) determines the topic assignments for the observed words by employing the conditional probability of the latent variables, where "$-ij$" indicates counts or variables with $z_{ij}$ excluded [50]. This specific conditional probability is defined as follows [51]:

$$p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon) = \frac{p(z_{ij}, z^{-ij}, X, w|\Omega, \Upsilon)}{p(z^{-ij}, X, w|\Omega, \Upsilon)} \tag{64}$$

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

$$p(z_{ij} = k|X, w, \Omega, \Upsilon) = \mathrm{E}_{p(z^{-ij}|w,X,\Omega,\Upsilon)}[p(z_{ij} = k|z^{-ij}, X, w, \Omega, \Upsilon)] \tag{65}$$

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 Equation (63) to be estimated.

$$p(X, z|w, \upsilon) = C \prod_{j=1}^{M} \left[ \frac{\Gamma(\sum_{i=1}^{k} \alpha_i) \Gamma(\alpha + \beta)}{\prod_{i=1}^{k} \Gamma(\alpha_i) \Gamma(\alpha) \Gamma(\beta)} \right] \times \frac{\prod_{i=1}^{k} \Gamma(\alpha_i^{'}) \Gamma(\alpha^{'}) \Gamma(\beta^{'})}{\Gamma(\alpha^{'} + \beta^{'}) \Gamma(\sum_{i=1}^{K} \alpha_i^{'})} \tag{66}$$

The likelihood of the multinomial distribution, defined by the parameter Y, and the probability density function of the Beta-Liouville distribution can be expressed as follows:

$$\begin{aligned} p(X|\Upsilon) &= \int p(X|\theta) p(\theta|\alpha_1, \ldots, \beta, \alpha) d\theta \\ &= \int \prod_{k=1}^{k} \theta_k^{m_k} \frac{\Gamma(\sum_{k=1}^{k} \alpha_k) \Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \prod_{k=1}^{K} \frac{\theta_k^{\alpha_k - 1}}{\Gamma(\alpha_k)} \\ &\times (\sum_{k=1}^{K} \theta_k)^{\alpha - \sum \alpha_k} (1 - \sum_{k=1}^{K} \theta_k)^{\beta - 1} d\theta \end{aligned} \tag{67}$$

By integrating the probability density function of the Beta-Liouville distribution over the parameter $\theta$ and incorporating updated parameters derived from the remaining integral in Equation (69), we are able to express it as a fraction of Gamma functions. The following shows the updated parameters, where $N_{jk}$ represents counts corresponding to variables [45,51]:

$$\begin{aligned} \alpha_K^{'} &= \alpha_k + \sum_{j=1}^{k} N_{jk} \\ \alpha^{'} &= \alpha + N_{jk} \\ \beta^{'} &= \beta + N_{jk} \end{aligned} \tag{68}$$

Equation (67) is then equivalent to

$$p(k|\alpha_1, \ldots, \alpha_k, \beta, \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{k=1}^{k-1} m_k)\Gamma(\beta+m_k)}{\Gamma(\alpha)\Gamma(\beta)\prod_{k=1}^{K}\Gamma(\alpha_k)\Gamma(\sum_{k=1}^{K}(\alpha_k+m_k))} \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+m_k)}{\Gamma(\alpha+\sum_{k=1}^{K-1} m_k+\beta+m_k)} \tag{69}$$

The parameters $\alpha_1, \ldots, \alpha_k, \alpha$, and $\beta$ 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 [47] is utilized to validate the results using a held-out dataset. Ultimately, this process leads to the estimation of the class conditional probability $p(X|w, \Omega, Y)$ within the framework of collapsed Gibbs sampling:

$$p(X|w, \Omega, Y) = \prod_{ij} \sum_{k=1}^{K} \frac{1}{S} \sum_{s=1}^{S} \tilde{\theta}_{jks} \tilde{\varphi}_{kws} \tag{70}$$

The parameters are then computed as follows:

$$\tilde{\theta}_{jks} = \frac{(N_{jk}+\alpha_k)(\alpha_{jk}+\sum_{l=k+1}^{K+1} N_{jl})(N_{jk}+\beta_k)}{(a_k b_k+\sum_{l=k+1}^{K+1} N_{jl})(\alpha_j+\sum_{l=k+1}^{K+1} N_{jl})} \tag{71}$$

$$\tilde{\varphi}_{kws} = \frac{(N_{jk}+\alpha_w)(\alpha_{jw}+\sum_{l=k+1}^{K+1} N_{jl})(N_{jk}+\beta_w)}{(\alpha_w b_w+\sum_{l=k+1}^{K+1} N_{jl})(\alpha_{wj}+\sum_{l=k+1}^{K+1} N_{jl})} \tag{72}$$

where S is the size of a sample.

## 4. Experimental Results

In this section, we validate our proposed algorithms' efficiency for two distinct and challenging applications, namely, topic modeling for medical text and sentiment analysis. Each model's evaluation is based on the success rate for each dataset and the perplexity [3,9,52,53], which is a common measure used in language modeling and is defined as:

$$prep(\mathcal{D}_{test}) = \exp\left(\frac{-\ln p(\mathcal{D}_{test})}{\sum_d |w_d|}\right) \tag{73}$$

where $|w_d|$ 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 analysis, 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.

### *4.1. Topic Modeling*

The goal of text classification is to assign documents to predefined subject categories, a problem extensively researched with various approaches [42,54,55]. Topic modeling, a common application in natural language processing, is used for analyzing texts from diverse sources and for document clustering [56]. 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 [12]. The "CMU Book Summary Dataset" was used to validate the model performance, containing plot summaries and metadata for 16,559 books [57]. The models' accuracy was tested by training on various document numbers and observing the impact of latent topics on the classification accuracy. Using variational Bayes inference, the models showed similar performances, but BLMPCA excelled, particularly in classifying similar classes.

In Tables 2–4, we present the first three topics, the perplexity measurements, and time complexity for 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 5 and 6, we show the results for the collapsed Gibbs sampling.

**Table 2.** Common topics identified with the BLMPCA model on the CMU Book dataset, each defined by a set of keywords.

| Topic Number | 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.** A comparison of the perplexity of the MPCA, GDMPCA, and BLMPCA models, indicating the 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 |

**Table 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 5.** Comparison perplexity scores of MPCA, GDMPCA, and BLMPCA, reflecting the model fit as the 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 |

**Table 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 |

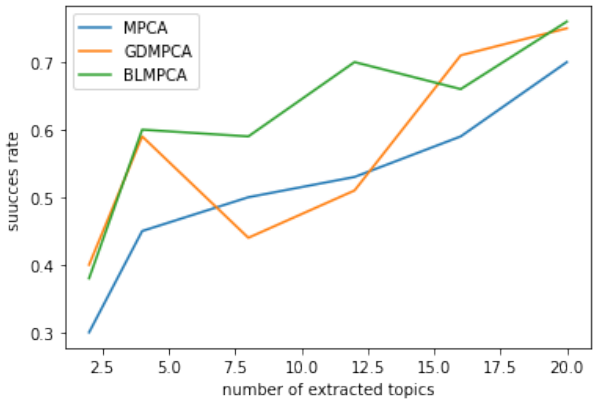


**Figure 1.** Success rate for CMU Book data.

### *4.2. 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. The study by Onan et al. [58] marked a significant advancement, presenting 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 utilized the TMVAr corpus from PubMed and the TMVAr-Dataset containing health-related Twitter news to evaluate models [59–64].

#### TMVAr Dataset

The TMVar Corpus dataset, comprising 500 PubMed papers with manual annotations of various mutation mentions, was utilized to evaluate our models. Tables 7 and 8 elucidate the perplexity comparison and time complexity for the TMVAR dataset, offering insight into the performances of our proposed methods. Moreover, Tables 9 and 10 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.

Furthermore, as shown in Table 11, the BLMPCA model successfully extracts pertinent topics, which is indicative of the model's nuanced analytical capabilities. Figure 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.

**Table 7.** Comparison of the perplexity of the MPCA, GDMPCA, and BLMPCA models, indicating the model fit quality across different topic numbers (K) on the 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 8.** Time complexity comparison for MPCA, GDMPCA, and BLMPCA with increasing topics (K) using variation EM inference on the 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 9.** Comparison of the perplexity of the MPCA, GDMPCA, and BLMPCA models, indicating the model fit quality across different topic numbers (K) on the 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 |

**Table 10.** 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 11.** Common topics identified with the BLMPCA model in the TMVAR dataset, each defined by a set of keywords.

| Topic Number | 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 |

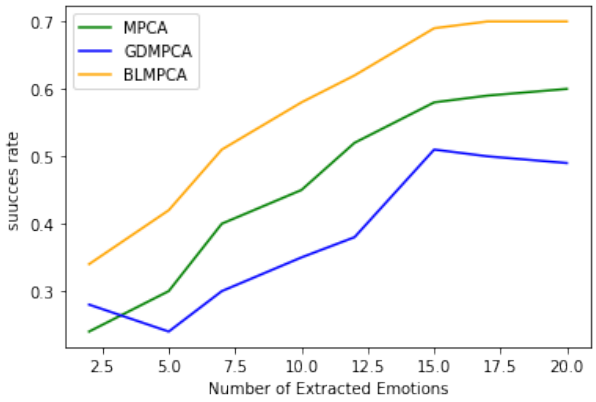


**Figure 2.** Success rate for Tmvar corpus data.

### *4.3. Sentiment Analysis*

Sentiment analysis, crucial for interpreting emotions in texts from various sources, benefits from advanced methodologies beyond mere word analysis [65]. Recent studies, such as [66,67], have demonstrated the effectiveness of deep learning and text mining in capturing nuanced sentiment expressions. Additionally, the authors of [68] highlighted the potential of ensemble classifiers in improving the sentiment classification accuracy. These innovations showcase the shift toward 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 [69]. This dataset, with extensive reviews on books and DVDs, provides data for basic analysis. The applied model, using K = 8 topics, assumed that each topic comprises a bag of words with specific probabilities, and each document is a mix of these topics. The model's goal was 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 12 and 13 provides 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 14 and 15. Figure 3 shows that our proposed models outperform the previous model. Figure 3 shows the success rates for MPCA, GDMPCA, and BLMPCA on sentiment analysis, with GDMPCA and BLMPCA outperforming MPCA as the number of emotions analyzed increases. This indicates their better suitability for complex emotion detection tasks in practical applications.

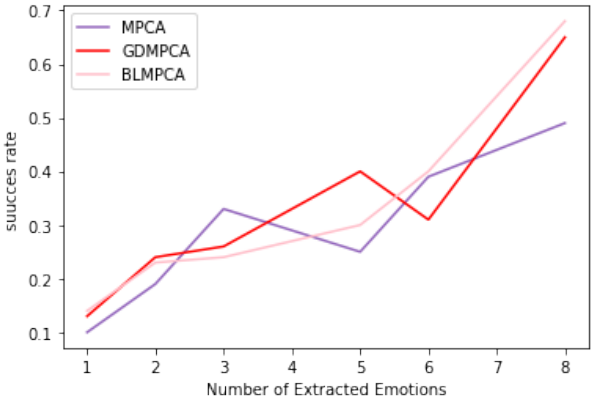


**Figure 3.** Success rate for sentiment dataset.

Furthermore, Tables 16 and 17 present the results for the collapsed Gibbs sampling. Additionally, Tables 18 and 19 display the accuracy and recall of various classifiers utilized for emotion detection. Table 20 shows the F1-scores for various classifiers, indicating the balanced harmonic mean of the precision and recall for SVM, Naive Bayes, and MLP classifiers when applied with MPCA, GDMPCA, and BLMPCA models in sentiment analysis.

**Table 12.** Comparison of the perplexity of the MPCA, GDMPCA, and BLMPCA models, indicating the model fit quality across different topic numbers (K) on sentiment data with variation EM inference.

| K | 2 | 3 | 5 | 8 |
|---|---|---|---|---|
| MPCA | 1551 | 1531 | 1542 | 1529 |
| GDMPCA | 1549 | 1539 | 1524 | 1521 |
| BLMPCA | 1448 | 1540 | 1531 | 1518 |

**Table 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 14.** Frequency of emotions identified in text data via topic modeling.

| Emotions | Count |
|---|---|
| satisfied | 78,901 |
| angry | 21,345 |
| happy | 6521 |
| joy | 82,345 |
| disgust | 7125 |
| Perfect | 45,459 |
| Tearful | 3451 |
| sad | 4387 |

**Table 15.** The counts of positive, negative, and unlabeled sentiments identified through sentiment analysis.

| Sentiment | Count |
|---|---|
| Positive | 213,232 |
| Negative | 36,308 |
| Unlabeled | 23,451 |

**Table 16.** Comparison of the perplexity of the MPCA, GDMPCA, and BLMPCA models, indicating the model fit quality across different topic numbers (K) on sentiment data with CGS inference.

| K | 2 | 3 | 5 | 8 |
|---|---|---|---|---|
| MPCA | 1451 | 1511 | 1589 | 1639 |
| GDMPCA | 1332 | 1393 | 1422 | 1502 |
| BLMPCA | 1316 | 1401 | 1413 | 1498 |

**Table 17.** 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 18.** 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 19.** 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 20.** 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 |

## 5. Discussion

We delved into the comparative advantages of the GDMPCA and BLMPCA models over existing methods in text classification and sentiment analysis. The superior performances of our proposed models can be attributed to several key factors. Firstly, the incorporation of Generalized Dirichlet and Beta-Liouville distributions allows for a more nuanced modeling of text data, which captures the intricacies of word distributions more effectively than traditional methods. This results in a more accurate representation of the underlying thematic structures in the data. For instance, in the CMU Book Summary Dataset, the intricacies of literary themes were better represented, showcasing the models' aptitude for multifaceted textual analysis. This was attributed to the models' ability to account for the co-occurrence and complex interrelationships of terms within documents, a feature less emphasized in MPCA due to its assumption of component independence.

In the TMVAR Corpus from PubMed, the medical text presented a challenge due to its specialized lexicon and the density of information. The BLMPCA model excelled by exploiting its additional parameters, optimizing data representation in this high-dimensional

space, thus underscoring the importance of model selection aligned with dataset characteristics. The sentiment analysis on the Multi-Domain Sentiment Dataset further emphasized the adaptability of our models. Here, BLMPCA demonstrated its finesse in discerning subtle sentiments from Amazon.com reviews, outperforming traditional approaches that may not have captured the emotional granularity present in user-generated content.

However, the sophistication of GDMPCA and BLMPCA comes with greater computational demands, as reflected in longer convergence times. This trade-off between accuracy and computational efficiency underscores the necessity of careful model selection in practice, considering the scale of the data and available computational resources. Although our proposed models signify a leap forward in text analysis methodologies, they are not without limitations. The reliance on variational inference and assumptions specific to the models may not be universally applicable to all types of textual data, suggesting room for future refinement and the exploration of alternative distributions or learning strategies.

The findings of this study illuminate the potential of integrating advanced probabilistic distributions into PCA to uncover deeper insights within text data. It is a testament to the evolution of statistical models in text analysis, pointing toward an exciting trajectory for future research in the field. The ongoing dialogue within the academic community on these topics is reflective of the dynamic nature of machine learning and its applications to natural language processing. As we continue to push the boundaries, it is imperative to balance innovation with practicality, ensuring that our models are not only theoretically robust but also computationally viable and accessible for varied applications.

## 6. Conclusions

In this paper, two novel models, generalized Dirichlet Multinomial PCA and Beta-Liouville Multinomial PCA, were proposed to improve the accuracy of the MPCA model for multi-topic modeling and text classification. We followed a Bayesian analysis that considers the generalized Dirichlet and Beta-Liouville assumptions. We demonstrated that our two proposed models have more flexibility. The models were used in two separate applications: text classification and sentiment analysis. The results show that the two proposed models, in all applications, achieved superior performances, as represented by the high prediction accuracy in comparison to that of the MPCA. It could be claimed that the proposed models, using different prior assumptions, yield better results than the standard methods. Specially, the BLMPCA provides the best improvement compared to the GDMPCA and MPCA for all the tested data. Crucially, the employment of collapsed Gibbs sampling for parameter inference was proven efficient and effective, despite its time-consuming nature. This method substantially boosts our models' computational capabilities, allowing for the superior discovery of latent topics in text corpora and marking a noteworthy advancement over the MPCA model. Future approaches for research will concentrate on model modifications and improvements to achieve greater precision in topic modeling. In addition, future works could be devoted to extending the proposed models to other applications and significantly extending the proposed model to fit a variety of data as well as real-time streaming data.

**Author Contributions:** P.K. was involved in the model's design methodology and implementation and writing the initial draft. E.I.K. helped with critical review, commentary, and revision. N.B. helped in critical review and revision, as well as with oversight and leadership responsibilities for the research activity planning. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** https://www.ncbi.nlm.nih.gov/research/bionlp/, https://www.cs.jhu.edu accessed on 3 October 2021

**Acknowledgments:** We would like to express our gratitude for the invaluable academic support received during the course of this research. Notably, we extend our thanks to the work of Shojaee Bakhtiari, Ali, whose dissertation "Count Data Modeling and Classification Using Statistical Hierar-

chical Approaches and Multi-topic Models", completed at Concordia University in 2014, provided foundational insights and methodologies that significantly guided our analysis and conclusions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| GD | Generalized Dirichlet; |
| BL | Beta-Liouville; |
| PCA | Principal Component Analysis; |
| LDA | Latent Dirichlet Allocation; |
| MPCA | Multinomial Principal Component Analysis; |
| PLSA | Probabilistic Latent Semantic Analysis; |
| pLSI | Probabilistic Latent Semantic Indexing; |
| SVD | Singular Value Decomposition; |
| NMF | Non-negative Matrix Factorization; |
| EM | Expectation-Maximization; |
| CGS | Collapsed Gibbs Sampling; |
| MCMC | Markov Chain Monte Carlo; |
| BLMPCA | Beta-Liouville Multinomial Principal Component Analysis; |
| GDMPCA | Generalized Dirichlet Multinomial Principal Component Analysis; |
| SVM | Support Vector Machine; |
| NLP | Natural Language Processing. |

## Appendix A. Exponential Family Distribution

The following introduces the general exponential family of distributions:

We have a vector of $T$ functions $t(x)$ and $d$ parameters $\theta$ for each individual sample point, which is a vector of measurements $x$, both of dimension $T$ and likely subject to some additional constraints. The following is the likelihood $q(x|\theta)$ [70]:

$$q(x|\theta) = \frac{1}{Y_t(x)Z_t(\theta)} exp(t(x)^{\mp}\theta) \tag{A1}$$

$Z_t(\theta)$ is modified to $Z$, or a distinguishing subscript is inserted. When $y$ is distributed as $q(y|\phi)$, the notation $E_{q(y|\phi)}$ is used to describe the expected value of quantity $A$. There are two main concepts that must be given [71]:

$$\begin{aligned} \mu_t &\equiv E_{q(y|\phi)}\{t(x)\} = \frac{\partial \log Z_t}{\partial \theta} \\ \Sigma_t &\equiv E_{q(y|\phi)}\{(t(x)-\mu_t)(t(x)-\mu_t)^{\mp})\} = \frac{\partial^2 \log Z_t}{\partial \theta \partial \theta} = \frac{\partial \mu_t}{\partial \theta} \end{aligned} \tag{A2}$$

The mean vector $\mu_t$ shares the same dimensionality as $\theta$, and the matrix $\Sigma_t$ encapsulates the covariance of $t(x)$, as noted in [20]. Notably, $\mu_t$ serves as a counterpart to the parameter set $\theta$. Specifically, when $\mu_t$ is fully ranked, it functions as the Hessian for changes in the basis. Moreover, $\mu_t$ represents the expected Fisher Information of the distribution. Both $t$ and $\Sigma_t$ can be directly derived from $Z_t$, indicating a unique relationship where $\mu_t$ acts as a complementary parameter set to $\theta$. In situations where $\mu_t$ 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, generalized Dirichlet, and Beta-Liouville distributions in Table A1. Another essential feature of the exponential family is the computation of the maximum a posteriori (MAP) estimates for parameters, derived from a dataset consisting of $I$ data points. 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 $\nu_t$ 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 [20].

$$\hat{\mu_t} = \frac{\nu_t + \Sigma_i t(x_i)}{S_t + I} \tag{A3}$$

**Table A1.** Exponential family characterizations for Dirichlet, GD and BL distributions.

| MODEL | $Z_t$ | $t_k(x)$ | $\theta_k$ | $\mu_{t,k}$ |
|---|---|---|---|---|
| Dirichlet | $\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$ | $\log(x_1), \ldots, \log(x_{k+1})$ | $\alpha_k$ | $\Psi_0(\alpha_k) - \Psi(\sum_k(\alpha_k))$ |
| GD | $\frac{\Gamma(a_i)\Gamma(b_i)}{\Gamma(a_i+b_i)}$ | $\log(x_1), \ldots, \log(1 - \sum_{t=1}^{D} x_t) - \log(1 - \sum_{t=1}^{D-1} x_t)$ | $a_k, b_k$ | $\Big(\Psi(a_i) - \Psi(a_i + b_i) + \sum_{m=1}^{i-1}(\Psi(b_m) - \Psi(a_m + b_m)\Big)$ |
| BL | $\frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ | $\log(x_1) - \log(\sum_{d=1}^{D} x_d) \ldots \log x_D \log(\sum_{d=1}^{D} x_d)$ | $\alpha_k, \alpha, \beta$ | $\Psi(\alpha) - \Psi(\alpha + \beta) + \Psi(\alpha_d) - \Psi(\sum_d \alpha_d)$ |

*Appendix A.1. The Generalized Dirichlet Distribution Exponential Form*

Since the GD distribution belongs to the exponential family of distributions, it can be represented in general as follows:

$$P(\theta|\xi) = Z_t(\theta) \times exp[\sum_{l=1}^{2d} G_l(\theta) T_l(\theta)] \tag{A4}$$

where

$$\begin{aligned}
Z_t(\theta) &= \prod_{l=1}^{d} \frac{\Gamma(\alpha_l + \beta_l)}{\Gamma(\alpha_l) + \Gamma(\beta_l)} \\
G_l(\xi) &= \alpha_l, (l : 1, \ldots, d) \\
G_l(\xi) &= \beta_{l-d} - \alpha_{l-d+1} - \beta_{l-d+1}, (l : d+1, \ldots 2d-1) \\
G_l(\xi) &= \beta_l (l : 2d) \\
T_l(\theta) &= \log(\theta_l), (l : 1, \ldots d) \\
T_l(\theta) &= \log(1 - \sum_{t=1}^{d-1} \theta_t), (l : d+1, \ldots 2d)
\end{aligned} \tag{A5}$$

In the formulation provided, $Z(\theta)$ represents the normalization factor, $G(\theta)$ is the natural parameter, and $T(\theta)$ denotes the sufficient statistics of the distribution. It is established that within the framework of the exponential family of distributions, the derivative of the logarithm of the normalization factor $Z(\theta)$ with respect to the natural parameters $G(\theta)$ is equal to the expected value of the sufficient statistics $T(\theta)$. This relationship underscores the fundamental connection between these components in statistical modeling within the exponential family. Therefore, we have

$$\begin{aligned}
E[\log(\theta_l)] &= \psi(\alpha_l + \beta_l) - \psi(\alpha - l) - \psi(\beta_l), l = 1, \ldots, d \\
E[\log(1 - \sum_{t=1}^{l} \theta_t)] &= \psi(\beta_l) - \psi(\alpha_l + \beta_l), l = 1, \ldots, d
\end{aligned} \tag{A6}$$

## Appendix B. Parameters for GDMPCA

In breaking down the L parameter for GDMPCA, we have the following.
By factorizing $\log p(w|\xi, \Omega) \geq E_q[(\theta, z, w)|\xi, \Omega] - E_q[\log q(z, \theta)]$, we have

$$\begin{aligned}\mathcal{L}(\gamma,\Phi;\xi,\Omega) = {} & E_q[\log p(m|\xi)] + E_q[\log p(c|m)] + E_q[\log p(w|c,\Omega)] \\ & - E_q[\log q(m)] - E_q[\log q(c)]\end{aligned} \tag{A7}$$

In the following, we derive each of the five factors of the above equation:

$$\begin{aligned}E_q[\log p(\theta|\xi)] = {} & \sum_{l=1}^{d}[\log\Gamma(\alpha_l+\beta_l) - \log\Gamma(\alpha_l) - \log\Gamma(\beta_l)] \\ & + \sum_{l=1}^{d}[\alpha_l(\Psi(\gamma_l) - \Psi(\gamma_l+\delta_l)) \\ & + (\Psi(\delta_l) - \Psi(\gamma_l+\delta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})]\end{aligned} \tag{A8}$$

$$E_q[\log p(z|\theta)] = \sum_{n=1}^{N}\sum_{l=1}^{d}\phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l+\delta_l)) + \sum_{n=1}^{N}\phi_{n(d+1)}(\Psi(\delta_d) - \Psi(\delta_d+\gamma_d)) \tag{A9}$$

$$E_q[\log p(w|z,\Omega)] = \sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{v}\phi_{nl}w_n^j\log(\Omega_{(lj)}) \tag{A10}$$

We should mention that $\Omega_{(lj)} = p(w_n^j = 1|z^l = 1)$:

$$\begin{aligned}E_q[\log q(\theta)] = {} & \sum_{l=1}^{d}(\log\Gamma(\gamma_l+\delta_l)\log\Gamma(\gamma_l) - \log\Gamma(\delta_l)) \\ & + \sum_{l=1}^{d}[\gamma_l(\Psi(\gamma_l) - \Psi(\gamma_l+\delta_l)) + (\Psi(\delta_l) - \Psi(\delta_l+\gamma_l)) \\ & (\delta_l - \gamma_{l+1} - \delta_{l+1})]\end{aligned} \tag{A11}$$

$$E_q[\log q(z)] = \sum_{n=1}^{N}\sum_{l=1}^{D+1}\phi_{nl}\log(\phi_{nl}) \tag{A12}$$

Subsequently, we will elaborate on Equation (A7) by expanding it with respect to both the model parameters and the variational parameters.

$$\begin{aligned}\mathcal{L}(\gamma,\Phi;\xi,\Omega) = {} & \sum_{l=1}^{d}[\log\Gamma(a_l+b_l) - \log\Gamma(a_l) - \log\Gamma(b_l)] \\ & + \sum_{l=1}^{d}[a_l(\Psi(\gamma_l) - \Psi(\gamma_l+\Phi)) \\ & + (\Psi(\Phi) - \Psi(\gamma_l+\Phi))(a_l - a_{l+1} - b_{l+1})] \\ & + \sum_{n=1}^{N}\sum_{l=1}^{d}m_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l+\Phi)) + \\ & \sum_{n=1}^{N}m_{n(d+1)}(\Psi(\Phi) - \Psi(\Phi+\gamma_d)) \\ & + \sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{v}m_{nl}w_n^j\log(\Omega_{ij}) \\ & - \sum_{l=1}^{d}(\log\Gamma(\gamma_l+\Phi)\log\Gamma(\gamma_l) - \log\Gamma(\Phi)) \\ & - \sum_{l=1}^{d}[\gamma_l(\Psi(\gamma_l) - \Psi(\gamma_l+\Phi)) + (\Psi(\Phi) - \Psi(\Phi+\gamma_l)) \\ & (\Phi - \gamma_{l+1} - \Phi_{l+1})]\end{aligned} \tag{A13}$$

*Appendix B.1. Variational Generalized Dirichlet*

To derive the update equations for the variational parameters in the generalized Dirichlet model, you start by isolating the terms in Equation (A7) that contain the variational parameters of the generalized Dirichlet. This involves examining the equation to identify which parts specifically involve these parameters and then focus on manipulating these parts to derive expressions for updating the parameters during the variational inference process. This method allows for the iterative refinement of the parameters, enhancing model accuracy with respect to the data being analyzed.

$$\begin{aligned} L[\xi_q] = & \sum_{l=1}^{d}[\alpha_l(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))(\beta_l - \alpha_{l+1} - \beta_{l+1})] \\ & + \sum_{n=1}^{N} \phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l) + \sum_{n=1}^{N} \phi_{n(d+1)}(\Psi(\gamma_d) - \Psi(\gamma_d + \delta_d)) \\ & - \sum_{l=1}^{d}(\log\Gamma(\gamma_l + \delta_l) - \log\Gamma(\gamma_l) - \log\Gamma(\delta_l)) + \sum_{l=1}^{d}(\Psi(\gamma_l) - \gamma_l(\Psi(\gamma_l + \delta_l)) \\ & + (\Psi(\delta_l) - \Psi(\delta_l + \gamma_l))(\delta_l - \gamma_{l+1} - \delta_{l+1}))) \end{aligned} \tag{A14}$$

Setting the derivative of the above equation to zero leads to the following updated parameters:

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \tag{A15}$$

$$\gamma_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \tag{A16}$$

Appendix B.1.1. Topic-Based Model

To derive the update equations for $\beta_w$, maximize Equation (A7) with respect to $\beta_w$. This involves setting the derivatives to zero, mirroring the optimization process used in MPCA, resulting in similar equations.

$$L[\beta_w] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{K+1} \sum_{j=1}^{V} \phi_{dnl} w_{dn}^{j} \log \beta_{w(lj)} + \sum_{l=1}^{K+1} \lambda_l \Big( \sum_{j=1}^{V} \beta_{w(ij)} \Big) \tag{A17}$$

Taking the derivative with respect to $\beta_{w(lj)}$ and setting it to zero yields

$$\beta_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \tag{A18}$$

In this scenario, because there are hidden variables present in the primary objective function, the situation is not fully addressed by Equations (33) and (34). However, the probability distribution $q(w|\gamma, r, m)$ can be accurately modeled using multinomials, which ensures that the minimum Kullback–Leibler (KL) divergence reaches zero. Consequently, the iterative updates will converge towars a local extremum of the log probability $\log p(\Omega, m|r)$.

$$\gamma_l = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \Omega m_{nl} \tag{A19}$$

$$m_{nl} = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \Omega_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi))} \tag{A20}$$

$$\Omega_{ij} = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} (2f_j + (\sum_{n} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi))}) \tag{A21}$$

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} m_{nl} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \Phi_l))} + m_{(d+1)n} e^{(\Psi(\Phi_d) - \Psi(\Phi_d + \gamma_d))}} \tag{A22}$$

Appendix B.1.2. Generalized Dirichlet Parameter

We select the components of Equation (A7) that involve the generalized Dirichlet parameters $\xi$.

$$\begin{aligned} L[\xi] = & \sum_{m=1}^{M} (\log(\Gamma(\alpha_l + \beta_l)) - \log\Gamma(\alpha_l)) - \log(\Gamma(\beta_l))) \\ & + \sum_{m=1}^{M} (\alpha_l(\Psi(\gamma_{ml} - \Psi(\gamma_{ml} + \delta_{ml})) + \beta_l(\Psi(\delta_{ml}) - \Psi(\delta_{ml} - \gamma_{ml}))) \end{aligned} \tag{A23}$$

Taking the derivative of the mentioned equation with respect to the generalized Dirichlet parameters yields

$$\frac{\partial L[\xi]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M} (\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{A24}$$

and

$$\frac{\partial L[\xi]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M} (\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{A25}$$

When applying the Newton–Raphson method to solve for the parameters, it is crucial to obtain the Hessian matrix with respect to the parameter space. The Hessian matrix of the likelihood function in this case assumes a particularly interesting form, as detailed below:

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\alpha_l)] \tag{A26}$$

$$\frac{\partial^2 L[\xi]}{\partial \beta_l^2} = M[\Psi^{'}(\alpha_l + \beta_l) - \Psi^{'}(\beta_l)] \tag{A27}$$

$$\frac{\partial^2 L[\xi]}{\partial \alpha_l \beta_l} = M[\Psi^{'}(\alpha_l + \beta_l)] \tag{A28}$$

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.

## Appendix C. Variational BLMPCA

To derive the parameter $\phi$, which represents the probability that the $n$-th word is generated by the $l$-th hidden topic, we maximize the relevant function with respect to $\phi$. This involves adjusting $\phi$ to optimize the likelihood of the observed data given the model's assumptions about topic distributions:

$$L[\phi_{nl}] = \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \phi_{ni} \log \beta_{w(iv)} - \phi_{ni} \log \phi_{ni} + \lambda_n(\sum_{l=1}^{D} \phi_{n(l)} - 1) \tag{A29}$$

and

$$\begin{aligned} L[\phi_{n(D+1)}] = & \phi_{n(D+1)}(\Psi(\beta_\gamma - \Psi(\alpha_\gamma + \beta_\gamma))) + \phi_{n(D+1)} \log \beta_{(D+1)v} \\ & - \phi_{n(D+1)} \log \phi_{n(D+1)} + \lambda_n(\sum_{i=1}^{D} \phi_{n(i)} - 1) \end{aligned} \tag{A30}$$

and therefore we have

$$\frac{\partial L}{\partial \phi_{nl}} = (\Psi(\gamma_d) - \Psi(\sum_{l=1}^{D} \gamma_l)) + \log \beta_{w(iv)} - \log \phi_{ni} - 1 + \lambda_n \tag{A31}$$

and

$$\frac{\partial L}{\partial \phi_{n(D+1)}} = (\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma)) \tag{A32}$$

Setting the above equation to zero leads to

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))} \tag{A33}$$

$$\phi_{n(D+1)} = \beta_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} \tag{A34}$$

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

$$e^{\lambda_n - 1} = \frac{1}{\beta_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} + \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))}} \tag{A35}$$

*Appendix C.1. Variational Beta-Liouville*

The updates mentioned were designed to converge to a local maximum of a lower bound of $\log p(\Omega, \mathrm{Y}|r)$, which is optimal for all product approximations such as $q(m)q(w)$ for the joint probability $p(m, w|\Omega, \mathrm{Y}, r)$. This approach ensures that the variational parameters are fine-tuned to best approximate the true posterior distributions within the constraints of the model.

$$\Phi_l = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)} m_{nl}(\lambda_n - 1)(\Psi(\gamma_l) - \Psi(\sum_{l=1}^{D} \gamma_l) \tag{A36}$$

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} m_{nl} \tag{A37}$$

$$\Omega_{(lj)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)} (2f \sum_{d=1}^{M} \sum_{n=1}^{N_d} m_{dnl} w_{dn}^{j}) \tag{A38}$$

In this case, variable $\Omega$ vanishes because m is defined in terms of the KL approximation. In the second step, the algorithm now optimizes for m. Since $q(w|\gamma, r, m)$ can be precisely modeled with multinomials, the minimum KL divergence is zero. As a result, the updates that follow converge to a local threshold of $\log p(\Omega, m|r)$:

$$\gamma_l = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)} \Omega m_{nl} \tag{A39}$$

$$m_{nl} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)} \Omega_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii})} \tag{A40}$$

$$\Omega_{ij} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)} (2f + (\sum_{n} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))}) \tag{A41}$$

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

$$e^{\lambda_n - 1} = \frac{1}{m_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} + m_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))}} \tag{A42}$$

## References


1. Aggarwal, C.C. An Introduction to Cluster Analysis. In *Data Clustering: Algorithms and Applications*; Aggarwal, C.C., Reddy, C.K., Eds.; CRC: Boca Raton, FL, USA, 2013; pp. 1–28.
2. Mao, J.; Jain, A.K. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Netw.* **1995**, *6*, 296–317. [PubMed]
3. Yu, S.; Yu, K.; Tresp, V.; Kriegel, H.P. A probabilistic clustering-projection model for discrete data. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 417–428.
4. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2267–2273.
5. Siddharthan, A.; Mani, I.; Maybury, M.T. (Eds.) *Advances in Automatic Text Summarization*; MIT Press: Cambridge, MA, USA, 1999.
6. Beeferman, D.; Berger, A.; Lafferty, J. Statistical models for text segmentation. *Mach. Learn.* **1999**, *34*, 177–210. [CrossRef]
7. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
8. Feldman, R. Techniques and applications for sentiment analysis. *Commun. ACM* **2013**, *56*, 82–89. [CrossRef]
9. Hua, T.; Lu, C.T.; Choo, J.; Reddy, C.K. Probabilistic topic modeling for comparative analysis of document collections. *ACM Trans. Knowl. Discov. Data (TKDD)* **2020**, *14*, 1–27. [CrossRef]
10. Cohn, D.A.; Hofmann, T. The missing link-a probabilistic model of document content and hypertext connectivity. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 430–436.
11. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
12. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
13. Ding, C.; He, X.; Zha, H.; Simon, H.D. Adaptive dimension reduction for clustering high dimensional data. In Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 9–12 December 2002; pp. 147–154.
14. Li, T.; Ma, S.; Ogihara, M. Document clustering via adaptive subspace iteration. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004; pp. 218–225.
15. Syed, S.; Spruit, M. Full-text or abstract examining topic coherence scores using latent dirichlet allocation. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 165–174.
16. Edison, H.; Carcel, H. Text data analysis using Latent Dirichlet Allocation: An application to FOMC transcripts. *Appl. Econ. Lett.* **2021**, *28*, 38–42. [CrossRef]
17. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
18. Collins, M.; Dasgupta, S.; Schapire, R.E. A Generalization of Principal Components Analysis to the Exponential Family. In Proceedings of the Advances in Neural Information Processing Systems 14: Natural and Synthetic, NIPS 2001, Vancouver, BC, Canada, 3–8 December 2001; pp. 617–624.
19. Buntine, W. Variational extensions to EM and multinomial PCA. In Proceedings of the European Conference on Machine Learning, Helsinki, Finland, 19–23 August 2002; pp. 23–34.
20. Jouvin, N.; Latouche, P.; Bouveyron, C.; Bataillon, G.; Livartowski, A. Clustering of count data through a mixture of multinomial PCA. *arXiv* **2019**, arXiv:1909.00721.
21. Griffiths, T.L.; Jordan, M.I.; Tenenbaum, J.B.; Blei, D.M. Hierarchical topic models and the nested chinese restaurant process. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 13–18 December 2004; pp. 17–24.
22. Hoffman, M.; Bach, F.R.; Blei, D.M. Online learning for latent dirichlet allocation. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–19 December 2010; pp. 856–864.
23. Fitzgerald, W.J. Markov chain Monte Carlo methods with applications to signal processing. *Signal Process.* **2001**, *81*, 3–18. [CrossRef]
24. Luo, Z.; Amayri, M.; Fan, W.; Bouguila, N. Cross-collection latent Beta-Liouville allocation model training with privacy protection and applications. *Appl. Intell.* **2023**, *53*, 17824–17848. [CrossRef] [PubMed]
25. Najar, F.; Bouguila, N. Sparse document analysis using beta-liouville naive bayes with vocabulary knowledge. In Proceedings of the Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, 5–10 September 2021; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2021; pp. 351–363.
26. Connor, R.J.; Mosimann, J.E. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.* **1969**, *64*, 194–206. [CrossRef]
27. Lacoste-Julien, S.; Sha, F.; Jordan, M.I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–10 December 2008; pp. 897–904.
28. Rabinovich, M.; Blei, D. The inverse regression topic model. In Proceedings of the International Conference on Machine Learning. PMLR, 2014, Beijing, China, 21–26 June 2014; pp. 199–207.

29. Ramage, D.; Hall, D.; Nallapati, R.; Manning, C.D. 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, Singapore, 6–7 August 2009; pp. 248–256.
30. Chemudugunta, C.; Smyth, P.; Steyvers, M. Modeling general and specific aspects of documents with a probabilistic topic model. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 241–248.
31. Ge, T.; Pei, W.; Ji, H.; Li, S.; Chang, B.; Sui, Z. 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), Beijing, China, 26–31 July 2015; pp. 575–585.
32. Onan, A. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access* **2019**, *7*, 145614–145633. [CrossRef]
33. Onan, A.; Toçoğlu, M.A. A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access* **2021**, *9*, 7701–7722. [CrossRef]
34. Meena, G.; Mohbey, K.K.; Indian, A.; Khan, M.Z.; Kumar, S. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimed. Tools Appl.* **2023**, *83*, 15711–15732. [CrossRef]
35. Meena, G.; Mohbey, K.K.; Kumar, S. Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100174. [CrossRef]
36. Meena, G.; Mohbey, K.K.; Kumar, S.; Lokesh, K. A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. *Decis. Anal. J.* **2023**, *7*, 100243. [CrossRef]
37. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
38. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [CrossRef]
39. Minka, T. *Estimating a Dirichlet Distribution*; Technical Report; MIT: Cambridge, MA, USA, 2003; Volume 1, p. 1.
40. Bouguila, N. Clustering of count data using generalized Dirichlet multinomial distributions. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 462–474. [CrossRef]
41. Bouguila, N.; Ziou, D. High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1716–1731. [CrossRef] [PubMed]
42. Bakhtiari, A.S.; Bouguila, N. A variational bayes model for count data learning and classification. *Eng. Appl. Artif. Intell.* **2014**, *35*, 176–186. [CrossRef]
43. Koochemeshkian, P.; Zamzami, N.; Bouguila, N. Flexible Distribution-Based Regression Models for Count Data: Application to Medical Diagnosis. *Cybern. Syst.* **2020**, *51*, 442–466. [CrossRef]
44. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [CrossRef]
45. Bouguila, N. Count Data Modeling and Classification Using Finite Mixtures of Distributions. *IEEE Trans. Neural Netw.* **2011**, *22*, 186–198. [CrossRef]
46. Ihou, K.E.; Bouguila, N.; Bouachir, W. Efficient integration of generative topic models into discriminative classifiers using robust probabilistic kernels. *Pattern Anal. Appl.* **2021**, *24*, 217–241. [CrossRef]
47. Espinosa, K.L.C.; Barajas, J.; Akella, R. The generalized dirichlet distribution in enhanced topic detection. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, 29 October–2 November 2012; pp. 773–782.
48. Shojaee Bakhtiari, A. Count Data Modeling and Classification Using Statistical Hierarchical Approaches and Multi-topic Models. Ph.D. Thesis, Concordia University, Montreal, QC, Canada, 2014.
49. Bakhtiari, A.S.; Bouguila, N. A latent Beta-Liouville allocation model. *Expert Syst. Appl.* **2016**, *45*, 260–272. [CrossRef]
50. Teh, Y.; Newman, D.; Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 1353–1360.
51. Ihou, K.E.; Bouguila, N. Stochastic topic models for large scale and nonstationary data. *Eng. Appl. Artif. Intell.* **2020**, *88*, 103364. [CrossRef]
52. Li, S.; Zhang, Y.; Pan, R. Bi-directional recurrent attentional topic model. *ACM Trans. Knowl. Discov. Data (TKDD)* **2020**, *14*, 1–30. [CrossRef]
53. Horgan, J. From complexity to perplexity. *Sci. Am.* **1995**, *272*, 104–109. [CrossRef]
54. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [CrossRef]
55. Riloff, E.; Lehnert, W. Information extraction as a basis for high-precision text classification. *ACM Trans. Inf. Syst.* **1994**, *12*, 296–333. [CrossRef]
56. Wallach, H.M. Topic modeling: Beyond bag-of-words. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 977–984.
57. Bamman, D.; Smith, N.A. New alignment methods for discriminative book summarization. *arXiv* **2013**, arXiv:1305.1319.
58. Onan, A. Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Comput. Math. Methods Med.* **2018**, *2018*, 2497471. [CrossRef]

59. Cohen, R.; Elhadad, M.; Elhadad, N. Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. *BMC Bioinform.* **2013**, *14*, 10. [CrossRef]
60. Wrenn, J.O.; Stein, D.M.; Bakken, S.; Stetson, P.D. Quantifying clinical narrative redundancy in an electronic health record. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 49–53. [CrossRef] [PubMed]
61. Karami, A.; Gangopadhyay, A.; Zhou, B.; Kharrazi, H. Flatm: A fuzzy logic approach topic model for medical documents. In Proceedings of the 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) Held Jointly with 2015 5th World Conference on Soft Computing (WConSC), Redmond, WA, USA, 17–19 August 2015; IEEE: Piscataway, NJ, USA; pp. 1–6.
62. Karami, A.; Gangopadhyay, A.; Zhou, B.; Kharrazi, H. A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections. In Proceedings of the iConference 2015, Newport Beach, CA, USA, 24–27 March 2015.
63. BIONLP. Available online: https://www.ncbi.nlm.nih.gov/research/bionlp/ (accessed on 3 October 2021).
64. Karami, A.; Gangopadhyay, A.; Zhou, B.; Kharrazi, H. Fuzzy approach topic discovery in health and medical corpora. *Int. J. Fuzzy Syst.* **2018**, *20*, 1334–1345. [CrossRef]
65. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R.J. Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011), Portland, OR, USA, 23 June 2011; pp. 30–38.
66. Onan, A. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e5909. [CrossRef]
67. Yan, X.; Li, G.; Li, Q.; Chen, J.; Chen, W.; Xia, F. Sentiment analysis on massive open online course evaluation. In Proceedings of the 2021 International Conference on Neuromorphic Computing (ICNC), Wuhan, China, 11–14 October 2021; pp. 245–249.
68. Onan, A.; Korukoğlu, S.; Bulut, H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Syst. Appl.* **2016**, *62*, 1–16. [CrossRef]
69. Blitzer, J.; Dredze, M.; Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czechia, 25–27 June 2007; pp. 440–447.
70. Gupta, R.D.; Kundu, D. Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biom. J. J. Math. Methods Biosci.* **2001**, *43*, 117–130. [CrossRef]
71. Buntine, W.L. Operations for learning with graphical models. *J. Artif. Intell. Res.* **1994**, *2*, 159–225. [CrossRef]