Bayesian Parameter Estimation of Probabilistic Models for Information Retrieval and Clustering in Discrete Data Spaces

Sahar Salmanzade Yazdi

A Thesis

in

The Department

of

Concordia Institue for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality System Engineering) at

Concordia University

Montréal, Québec, Canada

January 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Sahar Salmanzade Yazdi**

Entitled: **Bayesian Parameter Estimation of Probabilistic Models for Information Retrieval and Clustering in Discrete Data Spaces**

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

Master of Applied Science (Quality System Engineering)

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

Signed by the Final Examining Committee:

_____ Chair and Examiner
Dr. Arash Mohammadi

_____ Examiner
Dr. Walter Lucia

_____ Supervisor
Dr. Nizar Bouguila

Approved by _____
Dr. Chun Wang, Chair
Department of Concordia Institue for Information Systems Engineering

_____ 2024

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

Abstract

Bayesian Parameter Estimation of Probabilistic Models for Information Retrieval and Clustering in Discrete Data Spaces

Sahar Salmanzade Yazdi

In the contemporary era, a substantial amount of data is generated, prompting a critical need to effectively model data for thorough analysis and extraction of meaningful patterns. This is particularly crucial in various real-world applications, with natural language processing standing out as an area urgently requiring data analysis. Tasks such as document retrieval, spam email filtering, smart assistant applications, and sentiment analysis exemplify the extensive scope of natural language processing (NLP) and text mining. Addressing this context, various Bayesian models have been developed to aptly model data and extract essential information by considering latent topics. These models, grounded in probabilistic graphical models like Bayesian networks, capture the probabilistic dependencies between variables. Their ability to incorporate evidence from previous user knowledge enhances retrieval performance significantly. Furthermore, Bayesian network models exhibit effectiveness and generality surpassing classical information retrieval models like boolean, vector, and probabilistic models. This versatility positions Bayesian models as valuable approaches in information retrieval. Topic modeling, a valuable technique in text mining, plays a key role in uncovering concealed thematic structures within document collections, facilitates the identification of clusters of ”topics” or co-occurring words, and aids in understanding underlying themes and patterns from data. The unsupervised classification of documents, akin to clustering in numeric data, allows for the discovery of natural document groups, even in the absence of predefined topics.

However, significant challenges persist, including the management of queries not present in data collection, sparsity within datasets, especially in the age of big data, and addressing correlations between observations. This thesis suggests innovative Bayesian extensions for data modeling, utilizing

the Generalized Dirichlet distribution and the Beta-Liouville distribution as prior probability distributions to incorporate new queries into the topic space. Furthermore, these priors are integrated into a probabilistic clustering-projection model to evaluate their impact on both clustering and projection jointly. Lastly, in addressing the issues and hurdles associated with data sparsity, the Generalized Dirichlet distribution and the Beta-Liouville distribution are advocated as prior probability distributions to confront these challenges. The selection of a suitable prior is crucial in Bayesian data modeling, and these distributions are explored for their ability to model various non-Gaussian data and overcome the limited covariance structures of other distributions like the Dirichlet distribution. Following the determination of prior probabilities, the next step involves estimating optimized parameters for the distribution and model. An iterative parameter estimation model, utilizing the Expectation Maximization algorithm, is developed to maximize data likelihood. The simplicity of the proposed iterative algorithms allows these models to successfully handle real-time data, making them applicable across a broad range of practical scenarios.

Acknowledgments

I would like to express my heartfelt gratitude to Dr. Nizar Bouguila and Dr. Fatma Najar, for their exceptional patience, unwavering help, and steadfast support throughout my academic journey. Their guidance and mentorship have been invaluable, shaping my growth and contributing significantly to my academic accomplishments. My gratitude extends to my teachers, whose guidance propelled me forward in life, and now, I find myself deeply indebted to Dr. Bouguila for the academic standing I have achieved. It is a great honor to be his student on this transformative journey.

I extend sincere thanks to my parents, whose endless support has been the bedrock of my endeavors. Their encouragement, understanding, and belief in my abilities have been a constant source of motivation. I am deeply appreciative of the sacrifices they made to ensure my education and well-being.

I would also like to express my gratitude to my friends, whose encouragement and camaraderie have made the academic journey more enjoyable. Their support has been a source of strength during challenging times, and I am truly fortunate to have such a wonderful circle of friends.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the realm of Natural Language Processing (NLP), Information Retrieval involves the identification of relevant documents within unstructured datasets. This concept is not novel, tracing its roots back to 1949 post-World War II when American soldiers endeavored to decipher messages sent by Nazis. Inspired by this, Weaver (1952) proposed constructing a translation system. His idea centered on studying the common features present in the roots of all languages, positing that languages, being human creations with individuals sharing the same vocal organs and brain structure, would exhibit similarities. Weaver believed that "a book written in Chinese is simply a book that was written in English and later was coded into Chinese code". Over time, numerous scientists worked on automating the document search process. Luhn (1957) suggested employing a dictionary of 'notions', constructed by experts closely associated with the subject field. These notions, selected to optimally represent the material, would form the basis for automatically weighting them within documents. Luhn extended this concept to include the use of words themselves, aiming to extract vital sentences for automatic document abstracts Luhn (1958). While Luhn's systems were never built due to the technological constraints, they laid the groundwork for significant research ideas in subsequent decades. The 1970s and 1980s witnessed advancements building upon the groundwork laid in the 1960s. Various models of document retrieval emerged, demonstrating effectiveness on

small text collections. However, the scalability of these models to larger corpora remained uncertain.

In the late 20th century, with significant developments in the field of natural language processing and machine learning, the concept of topic modeling gained widespread interest and attention among researchers in various fields. The evolution of topic modeling is marked by key milestones, such as:

- **Latent Semantic Analysis (LSA)**: Proposed by Deerwester et al. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), Latent Semantic Analysis (LSA) laid the groundwork for topic modeling. LSA sought to unveil latent structures within large textual corpora by examining relationships between terms and documents.

- **Probabilistic Latent Semantic Analysis (pLSA)**: Serving as a precursor to LDA, Probabilistic Latent Semantic Analysis (pLSA) was proposed by Hofmann Hofmann (1999) as a statistical technique that models document generation in terms of topics, relying on a probabilistic generative model.

- **Latent Dirichlet Allocation (LDA)**: The introduction of Latent Dirichlet Allocation by Blei, Ng, and Jordan in 2003 D. M. Blei, Ng, and Jordan (2003) marked a significant advancement in topic modeling. LDA is a generative probabilistic model, explains sets of observations through unobserved groups, elucidating why certain data parts are similar. LDA has become widely employed in text mining and natural language processing.

The evolution of topic modeling has been characterized by the development of state of the art algorithms, the application of topic models to various types of texts, and the exploration of new research directions. Continuous efforts aim to enhance the capabilities and applications of topic modeling techniques. Bayesian models, known for their adept handling of uncertainty and integration of prior knowledge, are integral tools in text mining. Providing a probabilistic framework, they enable tasks like document classification, topic modeling, and sentiment analysis Ribeiro-Neto, Silva, and Muntz (2000). The role of prior knowledge is paramount in shaping topic and word distributions. In Bayesian topic modeling, the prior represents initial beliefs about topic and word

distributions before observing the actual data. Incorporating prior knowledge in Bayesian models offers several benefits, including:

- **Improved Model Performance**: Utilizing prior knowledge can boost the performance of topic models by guiding the inference process and aligning the model's output with existing domain knowledge. This guidance can result in more precise and understandable topic assignments, ultimately enhancing the overall quality of the model Yang, Downey, and Boyd-Graber (2015).

- **Enhanced Topic Coherence**: The integration of prior knowledge into topic models can generate more dependable and semantically meaningful topics that line up with human understanding and domain-specific concepts. This integration promotes results that are not only more interpretable but also more practical Wood, Tan, Wang, and Arnold (2017).

- **Effective Handling of Sparse or Noisy Data**: Prior knowledge serves as a valuable tool in addressing the challenges posed by sparse or noisy data. It provides additional constraints and guidance during the topic modeling process, leading to more resilient and dependable outcomes, particularly in situations with limited data Nguyen, Nguyen, Nguyen, Than, et al. (2021).

- **Facilitation of Interactive Topic Modeling**: The incorporation of prior knowledge facilitates interactive topic modeling by reducing the waiting time for model updates and enhancing the overall user experience. This improvement enables more efficient and effective collaboration between human experts and the topic modeling process Yang et al. (2015).

- **Improved Hierarchical Topic Discovery**: Using prior knowledge can elevate hierarchical topic modeling, resulting in the identification of topic hierarchies organized into interpretable order. This approach yields more structured and insightful representations of topics within the data Wang et al. (2022).

Bayesian method was first proposed by Bayes and Laplace in 18th century. The Baye's Theorem is defined as:

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \tag{1}$$

where $p(X|Y)$ is the posterior probability, and $p(Y|X)$ and $p(X)$ express data likelihood and prior probability, respectively Berrar (2018).

One of the most common used prior distribution is the Dirichlet distribution, which is a beneficial prior in Bayesian topic modeling owing to numerous advantages:

- **Conjugate Prior**: The Dirichlet distribution is the conjugate prior of the categorical and multinomial distributions. This means that when the prior distribution of the multinomial parameters follows the Dirichlet distribution, the posterior is also a Dirichlet distribution. This characteristic simplifies the computation of the posterior distribution Ferguson (1973).

- **Mathematical Convenience**: Using the Dirichlet distribution as a prior makes the mathematical calculations more straightforward and also simplifies the inference process and facilitates the computation of the posterior distribution Smucker and Allan (2005).

- **Modeling Uncertainty**: The Dirichlet distribution can be used to model the uncertainty of a random vector of probabilities, making it suitable for representing the distribution of topics within documents and the distribution of words within topics Hinneburg, Gabriel, and Gohr (2007).

- **Applications in Bayesian Inference**: Dirichlet distributions are commonly used as prior distributions in Bayesian inference, since they provide a flexible framework for representing prior beliefs and incorporating prior knowledge into the modeling process Gopalan and Berry (1998).

Let $X = (x_1, x_2, ..., x_K)$ and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$ be vectors of K parameters where for $i = 1, ..., K$, $x_i \geq 0$, $\alpha_i > 0$ and $\sum_{i=1}^{K} x_i = 1$. The probability density function (pdf) of the Dirichlet distribution is given by:

$$p(X|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{(\alpha_i - 1)} \tag{2}$$

where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. The mean, the variance, and the covariance of the Dirichlet distribution Şahin, Evren, Tuna, Şahinbaşoğlu, and Ustaoğlu (2023), for $i = 1, ..., K$, are as follows:

$$E(x_i) = \frac{\alpha_i}{\alpha_0} \tag{3}$$

$$Var(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{4}$$

$$Cov(\alpha_i \alpha_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{5}$$

Nevertheless, despite the advantageous characteristics of Dirichlet distributions, there are challenges associated with working with them. Specifically, when employing the Dirichlet distribution as a prior for the multinomial distribution, certain limitations emerge. All entries in the random vector must share a common variance, and their sum must equate to one. This restricts our ability to incorporate individual variance information for each entry of the random vector, as we are confined to a single degree of freedom, the total prior sample size, to express our confidence in prior knowledge. Additionally, all entries within the distribution are always negatively correlated. In simpler terms, if the probability of one entry increases, the probabilities of the remaining entries must either decrease or remain unchanged to ensure the total sums to one.

1.2 Contributions

The primary goal of this thesis is to assess the efficacy of employing alternative distributions, such as the generalized Dirichlet distribution and the Beta-Liouville distribution, as priors in recursive parameter estimation within latent topic modeling methods. This is pursued due to the enhanced flexibility these distributions offer in modeling diverse sets of data.

In this context, the contributions are listed as follows:

- **Bayesian Folding-In Using Generalized Dirichlet and Beta-Liouville Kernels for Information Retrieval** We propose a novel approach using generalized Dirichlet (GD) distribution and Beta-Liouville (BL) distribution as kernel densities in a Bayesian framework to involve the topic mixtures of the known documents and accurate modeling of these topic mixtures.

This contribution has been published in *2022 IEEE Symposium Series on Computational Intelligence (IEEE SSCI))* Yazdi, Najar, and Bouguila (2022)

- **Generalized Probabilistic Clustering Projection Models for Discrete Data** We extend our previous work by employing the generalized Dirichlet distribution and the Beta-Liouville distribution as priors to study the mutual effect of clustering and projection in a probabilistic framework and the impacts of prior knowledge in the perplexity of the model. This contribution has been published in *2023 IEEE International Symposium on Networks, Computers and Communications (ISNCC'23)* Yazdi, Najar, and Bouguila (2023)

- **Generalizing Conditional Naive Bayes model** We build upon our prior research efforts by expanding the scope of the conditional Naive Bayes model, incorporating the generalized Dirichlet and Beta-Liouville distributions in two distinct sets of experiments. In the first experiment, tailored for real-valued data, we adopt a Gaussian distribution to model the observed features. In the second experiment, designed for a categorical dataset, we consider discrete distributions to characterize the features. This contribution has been submitted to *2024 International Conference on Enterprise Information Systems (ICEIS)* Yazdi, Najar, and Bouguila (n.d.)

1.3 Thesis Overview

Each chapter provides a detailed explanation of the models we have developed.

- Chapter 2 presents an introduction to our Bayesian folding-in model, incorporating the generalized Dirichlet distribution and the Beta-Liouville distribution as priors. This model aims to determine the topic mixture proportions of new queries and capture dependencies between these topic mixtures and the topic mixture of the corpus. We illustrate the effectiveness of our model by comparing it with Bayesian folding-in using the Dirichlet distribution as a prior, utilizing real-life datasets.

- Chapter 3 delves deeper into examining the impact of the prior distribution on the perplexity of the probabilistic clustering-projection model. We conducted a comparison between the

Dirichlet distribution, the generalized Dirichlet, and the Beta-Liouville distributions using two datasets: the Reuters financial newswire service dataset and the 20 newsgroup data collection.

- Chapter 4 takes a different approach by focusing on a subset of features within the data collection, conditioning a Naive Bayes model. By employing exponential family distribution, we investigate the perplexity of the model under two scenarios: the generalized Dirichlet prior and the Beta-Liouville prior. This analysis utilizes data collections with both real-valued features and discrete features.

- In Chapter 5, we summarize all of our contributions and engage in a discussion to make a conclusion.

Chapter 2

Bayesian Folding-In Using Generalized Dirichlet and Beta-Liouville Kernels for Information Retrieval

2.1 Introduction

With the development of digital databases and networks, a large public would have access to vast libraries of textual data. Creating intelligent interfaces for human-machine interaction that assist users in their search for pertinent information is one of the biggest challenges facing the information science today. It is important to train a model which can distinguish between what users may say and what their actual intention is, especially since a user's request can be ambivalent, imprecise, or even unclear. To achieve that aim, latent semantic indexing (LSI) was proposed by Deerwester et al. Deerwester et al. (1990) to map terms as well as documents in the latent semantic space via singular value decomposition (SVD). However, LSI was expensive to compute, and it could not solve the polysemy problem. Therefore, probabilistic latent semantic indexing (PLSI) Hofmann (1999) was introduced to define a proper generative model of the data. PLSI has been presented to automatically index textual documents by representing them as mixtures of latent topics since it has a stronger statistical base than the traditional LSI. Yet, new queries had to be folded into the

latent topic space. PLSI folding-in and Bayesian folding-in were introduced to get the topic mixture proportions of new queries or documents that were not included in the document collection Brants, Chen, and Tsochantaridis (2002); Hinneburg et al. (2007). Studies have shown that topic mixtures for a new query could be ignored in PLSI folding-in which leads into a poor model of extended documents Hinneburg et al. (2007), whereas using Bayesian framework and maximum a posteriori estimate can express the correlation between topics. Furthermore, dependencies between topics also could be captured by modeling the prior using Dirichlet kernel in Bayesian framework. Although results have shown an affection in both precision and recall Hinneburg et al. (2007), the restrictive negative covariance structure of Dirichlet distribution makes its use as a prior in case of positively correlated data inappropriate. Thus, in this study, we propose using alternative priors such as generalized Dirichlet and Beta-Liouville kernels for accurate modeling of these topic mixtures. Although generalized Dirichlet distribution has a more general covariance structure than Dirichlet distribution Bakhtiari and Bouguila (2011), it still needs more parameters to be learned (i.e. a D-variate GD distribution has 2D parameters, while D+2 parameters for a D-variate Dirichlet distribution) Bouguila and Ziou (2009). Similar to generalized Dirichlet, Beta-Liouville distribution has a general covariance structure that can be either positive or negative. However, unlike generalized Dirichlet distribution, Beta-Liouville requires fewer parameters (i.e. a D-variate BL distribution has D+2 parameters), which considered as an important property of this distribution Najar and Bouguila (2021b).

In section 2 we review the PLSI and PLSI folding-in framework. Section 3 consists of Bayesian framework for folding new queries into latent topic space. The differences between Dirichlet, generalized Dirichlet, and Beta-Liouville kernels are examined in this section. In section 4, we describe our experiments on real text data and discuss the outcomes, and finally, in section 5 we make a conclusion.

2.2 PLSI FOLDING-IN

2.2.1 Probabilistic Latent Semantic Indexing (PLSI)

To overcome the problems with LSI, a statistical model was proposed by Hoffman as probabilistic latent semantic indexing (PLSI), which is based on likelihood principle Hofmann (1999). This way, it is possible to deal with polysemous words and distinguish between meanings and word usage patterns. A statistical model known as the aspect model serves as the foundation of PLSI Hofmann, Puzicha, and Jordan (1998); Saul and Pereira (1997). In aspect model, an unobserved class variable $z \in Z = \{z_1, z_2, \ldots, z_K\}$ is associated with each occurrence of the word $w \in W = \{w_1, w_2, \ldots, w_M\}$ in a document $d \in D = \{d_1, d_2, \ldots, d_N\}$. The generative model is defined by selecting document *d* with probability $P(d)$, while the latent class *z* with probability $P(z|d)$ is picked to produce a word *w* with probability $P(w|z)$. Respectively, the latent class variable *z* is ignored and the observation pair *(d, w)* is gained. To maximize the log-likelihood of occurring word *w* in document *d* with respect to *z*, the Expectation-Maximization (EM) algorithm is used. First, in the E-step, the probability that a word *w* in document *d* is explained by hidden factor *z*:

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')} \tag{6}$$

Then, the obtained probability is used to update the parameters of the model in the M-step.

$$P(w|z) = \frac{\sum_{d'} n(d,w)P(z|d,w)}{\sum_{d,n'} n(d,w')P(z|d,w')} \tag{7}$$

$$P(d|z) = \frac{\sum_{w'} n(d,w)P(z|d,w)}{\sum_{d',n} n(d',w)P(z|d',w)} \tag{8}$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w) \tag{9}$$

where,

$$R = \sum_{d,w} n(d,w) \tag{10}$$

n(d,w) indicates the number of occurrence of the word *w* in document *d* Hofmann (1999); Jin, Zhou,

and Mobasher (2004).

2.2.2 PLSI Folding-in

PLSI is a more adaptable method than LSI, therefore, it can be used in a range of information retrieval applications, text mining, and web usage mining. However, PLSI is not a generative model Hinneburg et al. (2007) and we cannot take advantage of it in case of new documents (queries). Therefore, a new document must be folded into the data collection through a process known as folding-in in order to determine the topic mixture proportions for that document.

The PLSI folding-in approach is similar to PLSI. However, in folding-in the obtained topic-word associations $P(w|z)$ from PLSI will be used to estimate the probability that a new query is described by the topic z, $P(z|d_q)$. Hence, by running the original EM algorithm with fixed topic-word associations $P(w|z)$, the EM algorithm turns down into:

$$\text{E-step} : P(z_j|w_i, d_q) = \frac{P(w_i|z_j)P(z_j|d_q)}{\sum_{j\prime=1}^{K} P(w_i|z_{j\prime})P(z_{j\prime}|d_q)} \tag{11}$$

$$\text{M-step} : P(z_j|d_q) = \frac{\sum_{i=1}^{V} n(d_q, w_i)P(z_j|w_i, d_q)}{n(d_q)} \tag{12}$$

The term $P(z_j|w_i, d_q)$ is calculated in E-step in each iteration using the updated $P(z_j|d_q)$ calculated from M-step Hinneburg et al. (2007).

Although the major drawback of PLSI folding-in is that in terms of short queries it is not able to find all latent aspects and that results in poor estimation and retrieving of relevant documents. As a consequence, a Bayesian way to estimate those topic mixtures for new queries (documents) was suggested, which instead of maximizing the likelihood, it maximizes the posterior $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$.

2.3 Bayesian Folding-in Using Dirichlet Kernel as Prior - Using EM Algorithm

Topic mixtures of documents in the collection could have been ignored in the former PLSI folding-in process, which can lead to inferior model of the extended collection. Folding-in technique in a Bayesian framework was introduced to eliminate such problems. In this method, topic mixtures

proportions $\vec{\theta_q}$ for a new query (document) will be estimated in a Bayesian way through maximizing the posterior probability $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$ while the prior is modeled using kernel density estimate. An alternative method for estimating the probability density function of a random variable is kernel density estimation (KDE), which is non-parametric. Kernel density estimation is a fundamental data smoothing problem where conclusions about the population are drawn from a small data sample. A kernel density estimate of a univariate probability density function f of independent and identically distributed sample of size N, $(X_1, X_2, \ldots, X_N)$, would be as follows:

$$\hat{f}_h(x) = N^{-1} \sum_{i=1}^{N} h^{-1} K(\frac{x - x_i}{h}) \tag{13}$$

where h is a smoothing parameter called bandwidth and K is a non-negative kernel Parzen (1962); Rosenblatt (1956).

In this method, a maximum a posteriori (MAP) estimator is used instead of maximum likelihood. The MAP approach is a form of Bayes estimate of a vector parameter where there is no information about the loss function. It provides a way to make use of prior knowledge in the training phase. This is especially useful to deal with the inaccuracies caused by sparse training data in maximum likelihood (ML) approach Bouguila (2009). The main difference between MAP and ML estimation is in the proposition of the suitable prior distribution of the parameter which is going to be estimated Gauvain and Lee (1994). For this MAP estimator, a kernel density estimate is employed using a Diriclet kernel as a prior. This kernel density estimate has the benefit of requiring very few model assumptions. Moreover, similar to the correlated topic model, such a prior can reflect correlations between topics.

In the study conducted by A. Hinneburg, et al. Hinneburg et al. (2007), they proposed using a kernel density estimate with Dirichlet distribution as the kernel function in MAP estimate to capture the relationships between topic mixtures of documents in the collection and topic mixtures of the new query. Since the Dirichlet distribution is a conjugate prior to the multinomial Bouguila (2007, 2009); Bouguila and Ziou (2004a), it is widely used as a prior in Bayesian statistics.

According to the definition of Dirichlet distribution Fang, Kotz, and Ng (2018), $(y_1, ..., y_k)$ is said to follow a Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_k)$:

$$(y_1, ..., y_K) \sim \text{Dir}(\alpha_1, ..., \alpha_K) \tag{14}$$

where:

$$\forall y_i : y_i \text{ is non-negative and } \sum_{i=1}^{K} y_i = 1$$

Therefore, the probability density function follows:

$$f(y_1, .., y_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} y_i^{\alpha_i - 1} \tag{15}$$

Although Dirichlet distribution is unimodal and assumes independence between topics, using it as the kernel can be multimodal and measures the dependencies between topics Hinneburg et al. (2007), which makes it more useful for the purpose of the study.

To derive the topic mixtures $\vec{\theta_q}$ of the new document d_q which maximizes $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$, the following three parameters are needed:

- $\vec{w_q}$: word vector for new document (query)
- $\vec{\theta}$: topic mixtures of the documents in the collection (training set)
- $\vec{w}$: topic-word associations.

Equation 16 shows posterior is related to the product of word likelihood $P(\vec{w_q}|\vec{\theta_q}, \vec{\theta}, \vec{w})$ and topic mixture prior $P(\vec{\theta_q}|\vec{\theta}, \vec{w})$.

$$P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w}) \propto P(\vec{w_q}|\vec{\theta_q}, \vec{\theta}, \vec{w}) P(\vec{\theta_q}|\vec{\theta}, \vec{w}) \tag{16}$$

First, considering the righthand side of (16), because of the independency between the word likelihood of the query and topic mixtures of the documents in the collection $(\vec{\theta})$, and also the basic assumption that all query words are independent, the word likelihood can be written as follow:

$$P(\vec{w_q}|\vec{\theta_q}, \vec{\theta}, \vec{w}) = \prod_{i=1}^{M} \sum_{j=1}^{K} P(w_i|a_j) P(a_j|d_q) = \prod_{i=1}^{M} \sum_{j=1}^{K} w_{ij} \theta_{qj} \tag{17}$$

Second, modeling prior by a kernel density estimate using Dirichlet kernels imposed two constraints mentioned in (14), which means topic mixtures are vectors with non-negative components

and the summation of all components of the vector is equal to one. To obtain the prior probability, this kernel density estimate sums over the topic mixtures of the documents in the training set. Since the prior does not depend on the word-topic association, we can write it as:

$$P(\vec{\theta_q}|\vec{\theta}) = \frac{1}{N}\sum_{l=1}^{N} \text{Dir}(\vec{\theta_q}|\alpha(\vec{\theta_l})) \tag{18}$$

Since it is difficult to maximize the righthand side of (16), two hidden variable vectors ($\vec{y}$and$\vec{z}$) are defined for word likelihood and prior distribution. It can be seen that the likelihood of a single word in a new document is a mixture model of *K* topics, so to show which topic a_j explains the word w_i, a hidden binary variable $\vec{y_i} \in \{0,1\}^K$ is defined. Similarly, $\vec{z} \in \{0,1\}^N$ is defined as a hidden binary variable to show the relationship between the Dirichlet component and the setting of the topic mixture of a new document.

According to the definitions and for the ease of calculations, the logarithm of the posterior will be maximized instead.

$$\begin{aligned}\log P(\vec{\theta_q}|\vec{w_q},\vec{y},\vec{z},\vec{w}) &= \log\left(P(\vec{w_q},\vec{y}|\vec{\theta_q},\vec{w})\,P(\vec{\theta_q},\vec{z}|\vec{\theta},\vec{w})\right) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{K} y_{ij}(\log w_{ij} + \log\theta_{qj}) + \sum_{l=1}^{N} z_l\left[\log\frac{1}{N} + \log Dir(\vec{\theta_q}|\alpha(\vec{\theta_l}))\right] + c\end{aligned} \tag{19}$$

Using EM algorithm, the probability for each hidden variable is calculated in the E-step and the topic mixture of the query is updated in the M-step. These two steps are done repeatedly to obtain the optimal topic mixture. Consequently, the maximum value for the posterior can be computed by putting the final value of topic mixtures in (19). The following equations show the E-step and M-step respectively Hinneburg et al. (2007).

E-step:

$$P(y_{ij}=1|w_i,\vec{\theta_q}^{(s)},\vec{w}) = \frac{w_{ij}\,.\,\theta_{qj}^{(s)}}{\sum_{j\prime=1}^{K} w_{ij\prime}\,.\,\theta_{qj\prime}^{(s)}} = g_{ij} \tag{20}$$

$$P(z_l=1|\vec{\theta_q}^{(s)},\vec{\theta}) = \frac{Dir(\vec{\theta_q^{(s)}}|\alpha(\vec{\theta_l}))}{\sum_{l\prime=1}^{N} Dir(\vec{\theta_q^{(s)}}|\alpha(\vec{\theta_{l\prime}}))} = h_l \tag{21}$$

M-step:

$$\theta_{qj}^{(s+1)} = \frac{\sum_{i=1}^{M} g_{ij} + \frac{1}{h}\sum_{l=1}^{N} h_l\theta_{lj}}{M+\frac{1}{h}} \tag{22}$$

2.4 Proposed Approaches

2.4.1 Bayesian Folding-In with Kernel Density Estimate using GD Kernel

In this part, we employ the kernel density estimate using a Generalized Dirichlet Distribution as the kernel.

Generalized Dirichlet (GD) Distribution

The GD was introduced to overcome the limitations of Dirichlet distribution as mentioned in Bouguila and ElGuebaly (2008a, 2008b); Bouguila and Ghimire (2010), particularly with modeling covariances, since using Dirichlet distribution as a prior requires random variables to be negatively correlated, where generalized Dirichlet does not have that restriction (as shown in Bakhtiari and Bouguila (2014)). Similar to the Dirichlet distribution, the generalized form of the Dirichlet distribution is also conjugate to multinomial distributionBouguila and Ghimire (2010); Najar and Bouguila (2022a); Wong (1998) however, its more general covariance structure not only makes it flexible and useful, especially in Bayesian analysis Bouguila (2011b); Connor and Mosimann (1969) but also provides high flexibility and simplicity of use for the approximation of both symmetric and asymmetric distributions Bouguila and Ziou (2007). The probability density function of the GD distribution is shown in equation (23). Note that, the random vector X has to be completely neutral. Connor and Mosimann Connor and Mosimann (1969) introduced the concept of neutrality as they faced the problem of analyzing the effects of eliminating one proportion, X_1, on the proportions $(X_2/[1-X_1], \ldots, X_k/[1-X_1])$. Assume $X = (X_1, X_2, \ldots, X_k)$ is a random vector of proportions. X_1 is neutral if it is independent of the vector $(X_2/[1-X_1], \ldots, X_k/[1-X_1])$ and has no effect on that. More generally, if we divide X into $(X_1, X_2, \ldots, X_j)$ and $(X_{j+1}, X_{j+2}, \ldots, X_k)$, X is completely neutral if for all $j < K$, $(X_1, X_2, \ldots, X_j)$ would be independent of $(X_{j+1}, X_{j+2}, \ldots, X_k)/(1-\sum_{i=1}^{j} X_i)$. Therefore, the density function for X following GD distribution with parameters $(\alpha_1, ..., \alpha_k; \beta_1, ..., \beta_k)$ in dimension k is given as follows:

$$f(x) = P(X_1, X_2, \ldots, X_k) = \prod_{i=1}^{k} \frac{1}{B(\alpha_i, \beta_i)} X_i^{\alpha_i - 1} (1 - \sum_{j=1}^{i} X_j)^{\gamma_i} \tag{23}$$

where, $B(\alpha_i, \beta_i)$ = $\Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i + \beta_i)$ and $\gamma_i = \beta_i - (\alpha_{i+1} + \beta_{i+1})$ for i = 1, ..., k-1, and $\gamma_k = \beta_k - 1$ Bouguila and ElGuebaly (2008b). For $\beta_i = \alpha_{i+1} + \beta_{i+1}$, GD distribution would be reduced to Dirichlet distribution. That indicates Dirichlet distribution is a special case of GD distribution Bouguila (2012); Bouguila and ElGuebaly (2009); Najar and Bouguila (2022b).

To use the generalized Dirichlet distribution instead of the Dirichlet distribution we need to redefine the prior as follow:

$$P(\vec{\theta_q}|\vec{\theta}) = \frac{1}{N}\sum_{l=1}^{N} GD(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\beta}) \tag{24}$$

Subsequently, the probability related to the specific setting of the topic mixture of the query changes. Since the probability of the hidden variable $\vec{y_i}$ is independent from the kernel estimate, it would remain as before. Therefore, the equations would be updates as follows:

E-step:

$$P(y_{ij} = 1|w_i, \vec{\theta_q}^{(s)}, \vec{w}) = \frac{w_{ij} \,.\, \theta_{qj}^{(s)}}{\sum_{j\prime=1}^{K} w_{ij\prime} \,.\, \theta_{qj\prime}^{(s)}} = g_{ij} \tag{25}$$

$$P(z_l = 1|\vec{\theta_q}^{(s)}, \vec{\theta}) = \frac{GD(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\beta})}{\sum_{l\prime=1}^{N} GD(\vec{\theta_q}|\alpha(\vec{\theta_{l\prime}}), \vec{\beta})} = f_l \tag{26}$$

M-step:

$$\theta_{qj}^{(s+1)} = \frac{\sum_{i=1}^{M} g_{ij} + \frac{1}{h}\sum_{l=1}^{N} f_l \theta_{lj}}{M + \frac{1}{h}} \tag{27}$$

By replacing the topic mixture obtained from the EM algorithm in (19) the logarithm of the posterior is given as follows:

$$\begin{aligned} \log(P(\vec{\theta_q}|\vec{w_q}, \vec{y}, \vec{z}, \vec{w})) &= \log\left(P(\vec{w_q}, \vec{y}|\vec{\theta_q}, \vec{w})\, P(\vec{\theta_q}, \vec{z}|\vec{\theta}, \vec{\alpha})\right) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{K} y_{ij}[\log w_{ij} + \log \theta_{qj}] + \sum_{l=1}^{N} z_l\Big[\log\frac{1}{N} + \log GD(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\beta})\Big] + c \end{aligned} \tag{28}$$

2.4.2 Bayesian Folding-In with Kernel Density Estimate using BL Kernel

In addition of using generalized Dirichlet as a prior in Bayesian framework, here we modeled the prior with a member of Liouville family of distributions called Beta-Liouville Fan and Bouguila (2015) distribution. Beta distribution was selected as a generating density because of its flexible

shape and ability to approximate nearly any arbitrary distribution Bouguila and Ziou (2007); Najar and Bouguila (2022b).

Beta-Liouville (BL) Distribution

Despite the fact that Dirichlet involves a small number of parameters (D+1 parameters for a D-variate Dirichlet), its negative covariate matrix makes its application restrictive in many areas, such as pattern recognition, data mining, and computer vision. Thus, the generalized Dirichlet distribution was proposed to overcome the problem of negative covariance matrix since it has a more general covariance structure, which includes positive and negative values. However, compared to Dirichlet, generalized Dirichlet has more parameters (2D parameters are needed to define a D-variate generalized Dirichlet). In this section, we use Beta-Liouville as a kernel since it has been demonstrated to be a good substitute for the Dirichlet and generalized Dirichlet distributions in statistically representing of proportional data Daghyani, Zamzami, and Bouguila (2019); Fan and Bouguila (2015) and an appropriate choice for classification in case of positive data Zamzami and Bouguila (2020a). Similar to generalized Dirichlet, Beta-Liouville has a general covariance structure but compared to GD, less parameters are needed to define Beta-Liouville (a D-variate Beta-Liouville is defined by D+2 parameters) Bouguila (2010, 2012). The probability density function of a D-dimensional vector $\vec{X} = (X_1, X_2, \ldots, X_D)$ following the Beta-Liouville distribution with positive parameters $(\alpha_1, \alpha_2, \ldots, \alpha_D, \alpha, \beta)$ is given by (29) while $\sum_{d=1}^{D} X_d < 1$ and $X_d > 0$ for $d = 1, ..., D$.

$$P(\vec{X}|\alpha_1, \alpha_2, ;\alpha_D, \alpha, \beta) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{X_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times (\sum_{d=1}^{D} X_d)^{\alpha - \sum_{d=1}^{D} \alpha_d} \left(1 - \sum_{d=1}^{D} X_d\right)^{\beta - 1} \tag{29}$$

Note that the Dirichlet is a special case of Beta-Liouville. This can be proved assuming that the density generator has a Beta distribution with parameters $\sum_{d=1}^{D} \alpha_d$ and α_{D+1} Bouguila (2010).

Equation (30) shows using Beta-Liouville distribution as a kernel in defining the prior. Again, because there is no dependency between the hidden variable $\vec{y_i}$ and the kernel estimate, the calculation for $P(y_{ij} = 1|w_i, \vec{\theta_q}^{(s)}, \vec{w})$ would be the same as (20). However the value for $P(z_l = 1|\vec{\theta_q}^{(s)}, \vec{\theta})$ needs to be updated for this part (equation 31).

$$P(\vec{\theta_q}|\vec{\theta}) = \frac{1}{N}\sum_{l=1}^{N} BL(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\alpha}, \vec{\beta}) \tag{30}$$

$$P(z_l = 1|\vec{\theta_q}^{(s)}, \vec{\theta}) = \frac{BL(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\alpha}, \vec{\beta})}{\sum_{l\prime=1}^{N} BL(\vec{\theta_q}|\alpha(\vec{\theta_{l\prime}}), \vec{\alpha}, \vec{\beta})} = t_l \tag{31}$$

By replacing the value obtained from (31) in (19) the topic mixture of the query would be calculated in M-step. The resulted topic mixture of the query would be used to determine the maximum value of the logarithm of the posterior (33).

$$\theta_{qj}^{(s+1)} = \frac{\sum_{i=1}^{M} g_{ij} + \frac{1}{h}\sum_{l=1}^{N} t_l \theta_{lj}}{M + \frac{1}{h}} \tag{32}$$

$$\begin{aligned}\log[P(\vec{\theta_q}|\vec{w_q}, \vec{y}, \vec{z}, \vec{w})] &= \log\left(P(\vec{w_q}, \vec{y}|\vec{\theta_q}, \vec{w})P(\vec{\theta_q}, \vec{z}|\vec{\theta}, \vec{w})\right) \\ &= \left[\sum_{i=1}^{M}\sum_{j=1}^{K} y_{ij}[\log w_{ij} + \log \theta_{qj}]\right] \\ &+ \sum_{l=1}^{N} z_l \left[\log\frac{1}{N} + \log BL(\vec{\theta_q}|\alpha(\vec{\theta_l}), \vec{\alpha}, \vec{\beta})\right] + c\end{aligned} \tag{33}$$

The general approach to maximize a posterior is presented in Algorithm 1. Consider using GD distribution as a prior. After assigning initial values to the model parameters and the kernel parameters we begin by calculating the probability of each hidden variable g_{ig}, and f_l according to equations 25 and 26. In the next step, using the calculated variables, we update the topic mixtures ($\theta_{qj}^{(s+1)}$) using equation 27. Using those values, we calculate the logarithm of the posterior probability $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$ using equation 28. In this recursive classification algorithm, we receive the query data in real time and update the parameters g_{ig}, f_l, and $\theta_{qj}^{(s+1)}$, and repeat the process in an iteration loop until we meet the termination criterion (the difference between the new log-posterior and the old log-posterior in iteration t and $t-1$ is less than a threshold $\epsilon = 0.5$).

2.5 Experimental Results

The aim of this study is to find out how Bayesian folding-in framework using GD and BL kernels can affect retrieving information compared with using Dirichlet kernels as a prior in Bayesian

Algorithm 1 Maximum a posteriori algorithm with GD and BL kernels as prior

Result: Log-psteriori $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$
Input: Text document collection
1: **Initialization:** initialize model parameters $P(w_i|z_j)$ and $P(z_j|d_q)$, ($\vec{\alpha}$, $\vec{\beta}$) for the GD prior and ($\vec{\alpha}$, α, β) for the BL prior.
2: **While** new Log Posteriori - old Log Posteriori $< \epsilon$ **do**
3: **E step:**
4: Calculate the values of g_{ij} and $P(z_l = 1|\theta_q^{\vec{(s)}}, \vec{\theta})$
5: **M step:**
6: Update the topic mixtures $\theta_{qj}^{(s+1)}$
7: Calculate the new log-posteriori $P(\vec{\theta_q}|\vec{w_q}, \vec{\theta}, \vec{w})$
8: **end While**

framework. To achieve that purpose, 3 text-based document collections are used, namely CISI (Centre for Inventions and Scientific Information), CRAN (aerodynamic, aeroelastic analysis), and MED (medical records)[1]. Each collection consists of the training documents, query documents, and the ground truth document. The ground truth document represents the queries with their relevant documents. To evaluate the performance and efficiency of methods that were mentioned in previous sections, precision and recall are defined as comparison parameters. For a given document collection and a query, precision and recall are defined in (34) and (35), where a is a set of relevant documents and b represents retrieved documents Hinneburg et al. (2007).

$$Precision : \frac{|a \cap b|}{|b|} \tag{34}$$

$$Recall : \frac{|a \cap b|}{|a|} \tag{35}$$

First step is preprocessing documents for further assessments. This includes elimination of stop words and infrequent words. The term infrequent refers to the words that occur in less than δ documents. In order to insure that no queries will be empty, δ is considered to be 5. Therefore, documents that have less than 5 words are neglected. Moreover, all terms are reduced to their stems using Porters stemmer.

After preprocessing, PLSI is performed on training documents for 32 topics to estimate a vector of learned document topic mixtures in the latent topic space. Then, queries are folded into obtained co-occurrence model through 4 approaches. The observed query-topic mixtures are used to

[1] http://ir.dcs.gla.ac.uk/resources/

determine the similarities between each query and the training documents using cosine similarity.

Table 2.1 shows the observed values of recall and precision after the 3rd iteration in each framework for the datasets.

Table 2.1: Recall and precision values observed for different Datasets

Dataset	Framework	Recall	Precision
CISI	PLSI	0.063925	0.020842
	Dir	0.625763	0.019796
	GD	0.858015	0.019333
	BL	0.758432	0.019270
CRAN	PLSI	0.112745	0.005896
	Dir	0.640523	0.005906
	GD	0.803922	0.005845
	BL	0.611112	0.005734
MED	PLSI	0.004316	0.008746
	Dir	0.303597	0.022579
	GD	0.766906	0.0225695
	BL	0.726619	0.022836

According to Fig. 2.1, for the first two methods, PLSI folding-in and Bayesian folding-in using Dirichlet kernel, the value for recall decreases after each iteration which indicates reduction in the ability of the method to find relevant documents and queries. Whereas for Bayesian folding-in using generalized Dirichlet kernel and Bayesian folding-in using Beta-Liouville kernel, the value for recall increases gradually after each iteration. Unlike stable changes in recall values, the value for precision changes quietly in every iteration because of the dependency between precision and retrieved documents. Moreover, results show that apparent constraints on covariances in GD and BL distributions result in better modeling of sparse topic correlations in natural language documents. Consequently, empirical likelihood and information retrieval metrics perform better using these priors.

2.6 Conclusion

The purpose of this study was to employ different techniques to extract the relevant documents given a new query. Generalized Dirichlet and Beta-Liouville distributions had been used as kernels in Bayesian framework to examine the performance of the method compared to PLSI folding-in

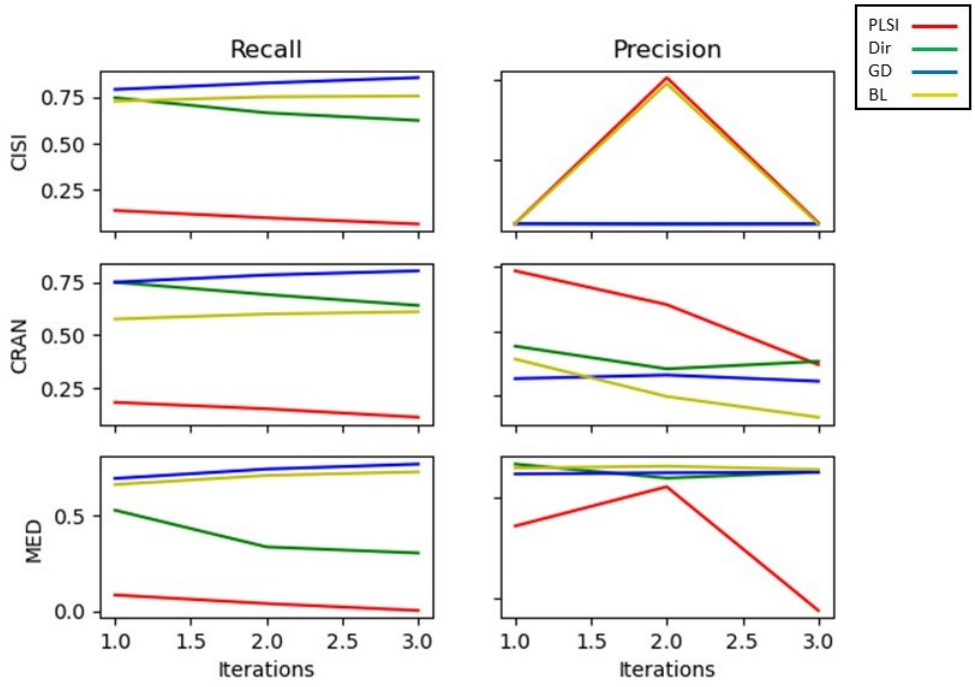


Figure 2.1: A comparison between recall and precision in different frameworks for three datasets.

and Bayesian folding-in using Dirichlet kernel. Generalized Dirichlet and Beta-Liouville had been proven to be more efficient due to their more general covariance structure while Dirichlet has restrictive negative covariance. Moreover, less variables are required to define Beta-Liouville distribution. Observations indicated an improvement in recall values for generalized Dirichlet and Beta-Liouville kernels compared to former techniques. Prospective studies could be devoted to apply different initialization methods or to employ other distributions as kernel in Bayesian framework in order to investigate the effect of the initial values and the kernel function on the accuracy of parameter estimation at the end.

Chapter 3

Generalized Probabilistic Clustering Projection Models for Discrete Data

3.1 Introduction

Information can be retrieved from numerous databases using data mining techniques. When it comes to textual data, data warehouses were found to be ineffective yet successful for numerical information. Unlike traditional information retrieval systems that required users to identify relationships within and across documents, methods such as GD-LDA Caballero, Barajas, and Akella (2012), Pachinko Allocation Model (PAM) W. Li and McCallum (2006), Dirichlet Process Mixture Model with Feature Partition (DPMFP) Huang, Yu, Wang, Zhang, and Shi (2012), and PCP model Yu, Yu, Tresp, and Kriegel (2005) have demonstrated increased efficiency and the ability to provide relevant results based on user requests. The process of extracting valuable information from text-based data which is called text mining, generally entails formatting the input text, looking for the patterns in formatted data, assessing, and then interpreting the results. Topic models are considered generative models and are typically used in natural language processing to automatically categorize, comprehend, search, and summarize large data archives. The term 'topic' refers to the unobserved relations between words in a vocabulary and their occurrence in data collection. Yet, finding out those topics in high-dimensional text data is challenging.

Different methods have been proposed to overcome that challenge. The first step is to transfer

the data into a lower-dimensional space. Projection techniques such as PCA Abdi and Williams (2010), and Linear discriminant analysis Xanthopoulos, Pardalos, and Trafalis (2013) help us to project the data into a lower dimension space by mapping the features to find a new representation of the data. Additionally, clustering methods such as k-means Hartigan and Wong (1979), PAM (K-Medoids) Reynolds, Richards, and Rayward-Smith (2004), and DBSCAN Ester, Kriegel, Sander, Xu, et al. (1996) are used to group similar documents based on their structural pattern. Several studies were conducted to model the connections between the documents based on the relationships between terms and their occurrence in documents D. M. Blei et al. (2003); Deerwester et al. (1990); Hofmann (1999); Yu et al. (2005). Latent semantic indexing (LSA) was introduced by Deerwester et al. Deerwester et al. (1990) as an unsupervised learning technique, to uncover similarities among documents by building a word-document matrix, in which rows are individual words, columns are documents, and each cell represents the word frequency associated with the appearance of the word in its row in the document denoted by its column. By applying singular value decomposition (SVD) on the matrix a K-dimensional vector (known as a topic vector) is derived which LSA uses for its semantic space. Finally, similarities are determined between entities in the reduced-dimensional space instead of the original term-document matrix Dumais et al. (2004). LSA has demonstrated its ability to simulate a wide range of human cognitive phenomena, including the acquisition of recognition vocabulary during development, word categorization, semantic priming between sentences and words, comprehension of discourse, and evaluations of essay quality, through its generation of word and passage meaning representations. However, LSA has some limitations. It does not take into account word order, syntactic relations, logic, or morphology. Surprisingly, despite the absence of these factors, LSA manages to capture passage and word meanings quite accurately. Nonetheless, there is still a possibility of occasional incompleteness or errors. Furthermore, the relationship between certain neighboring words in LSA space can be perplexing, with some expected pairs not appearing close together. The exact reasons behind these anomalies are difficult to determine, but it is plausible that words with multiple contextual meanings (polysemous words) are assigned an average placement in high-dimensional space, which may not convey any specific meaning when taken out of context. Additionally, many words may be inadequately represented due to insufficient sampling. It is also possible that LSA's reliance on the 'bag of words' approach, which disregards

syntactical, logical, and non-linguistic implications, occasionally results in missed or scrambled meaning Landauer, Foltz, and Laham (1998). Therefore, a probabilistic version of LSA called PLSA (probabilistic latent semantic analysis) was introduced by Hofmann Hofmann (1999) to define a generative model of the data. In PLSA, the conditional probability between documents d and words w is modeled through a latent variable z called topic. Hofmann showed that in addition to reducing word perplexity, PLSA can take advantage of statistical standard methods for model fitting, controlling overfitting, and combining models. His evaluations demonstrated that PLSA achieves significant improvements in precision compared to both standard term matching and LSA, which clearly confirmed the advantages of PLSA. Latent Dirichlet Allocation (LDA) D. M. Blei et al. (2003) was presented as an alternative approach to generating a probabilistic model of a corpus. Since LDA models documents in a Bayesian framework, it is more useful, especially because of using adjustable priors which are nearly similar to the structure of the documents, topics, and terms. LDA assumes that K latent topics are hidden in a collection of N documents. Also, each of those topics follows a multinomial distribution over words in the extracted vocabulary from the document collection. The final topic-word and document-topic mixtures are obtained by maximizing the posterior with Gibbs sampling Chen, Zhang, Liu, Ye, and Lin (2019). Although mentioned methods have been applied with remarkable success in different text-processing tasks, still the joint effects of projection and clustering are needed to be considered since both of them capture the inherent structure of the data and are expected to mutually strengthen each other. For that purpose, the probabilistic clustering-projection (PCP) model was introduced by Yu et al. Yu et al. (2005) to examine the mutual impact of projection and clustering for discrete data. In their model, a Dirichlet prior was put for topic mixtures over the topic space, which was denoted as a vector of cluster centers. However, due to the restrictive negative covariance structure of Dirichlet distribution, its applications as a prior in the case of positively correlated data would be limited. Thus, in this paper, we propose to use alternative distributions namely generalized Dirichlet (GD) distribution and Beta-Liouville (BL) distribution as priors in defining topic mixtures since due to their broader covariance structures, using them as a prior would lead to a more precise model.

The chapter is structured as follows, section 2 is a brief review of the PCP model. In section

3, we propose our alternative PCP model using GD and BL distributions and we discuss the differences them. Using real text documents, in section 4, we describe our experiments and interpret the outcomes. And finally, in section 5, we have the conclusion.

3.2 Probabilistic clustering-projection model

Probabilistic clustering-projection (PCP) model was proposed as a generative model to consider both projection and clustering together. In this model, β is defined as a $K \times V$ projection matrix specifying the probability of generating each word given the corresponding topic, while topic mixtures are presented by cluster centers in a K-dimensional topic space. A specific document-topic mixture θ_d is obtained after generating document d by choosing a cluster from M clusters which leads to defining an indicator value c_d to show that among M clusters which cluster document d would take on. Given those topic mixtures, each word is generated by choosing a topic from the topic mixtures and then sampling the word given β. The PCP model could be considered as a clustering model when the projection β is known. The probability that document d belongs to the cluster M is calculated by projecting the words into the topic space and measuring the distance of all words to the cluster center. In the PCP model, topics are modeled by cluster centers, so if the clustering structure is known, PCP can use the information in the cluster center to learn the projection β to map words to topics. Therefore, this way PCP could be explained as a projection model Yu et al. (2005).

Assuming that α, λ, and β are model parameters, and θ, π, and w_d denote cluster centers, mixing weights, and sequence of the words in document d respectively, the likelihood of the corpus $\mathcal{D}$ would be calculated as:

$$\mathcal{L}(\mathcal{D}; \alpha, \lambda, \beta) = \int_{\pi} \int_{\theta} \prod_{d=1}^{D} p(w_d|\theta, \pi; \beta)\ dP(\theta; \lambda)\ dP(\pi; \alpha) \tag{36}$$

However, since we are working with high dimensional discrete data (text documents) calculating the integrals would be intractable and highly cost time-wise. Thus, a variational EM algorithm has been derived which iteratively maximizes the lower bound of the log-likelihood. $q(\pi, \theta, c, z|\eta, \gamma, \psi, \phi)$ is proposed as the variational distribution and it is a function of the current mixture parameters and a variational distribution over latent variables.

$$q(\pi, \theta, c, z|\eta, \gamma, \psi, \phi) = q(\pi|\eta) \prod_{m=1}^{M} q(\theta_m|\gamma_m) \prod_{d=1}^{D} q(c_d|\psi_d) \prod_{n=1}^{N_d} q(z_{d,n}|\phi_{d,n}) \tag{37}$$

η, γ, ψ, and ϕ are representing the variational parameters describing an arbitrary distribution over model parameters:

- η: M-dim. Dirichlet for π
- γ: K-dim. Dirichlet over θ
- ψ: M-dim. multinomial for indicator value c_d
- ϕ: K-dim. multinomial over latent topics for word $w_{d,n}$

3.2.1 Variational EM Algorithm

Suppose that *y* is the observed data, and *x* and Θ are vectors of latent variables and parameters, respectively. To obtain the maximum likelihood estimate of Θ and the posterior distribution of *x* we have to calculate the maximum log-likelihood $\log p(y; \Theta)$. Maximum log-likelihood computation, however, is usually challenging. Using Baye's rule we can break down $p(y; \Theta)$ concerning the variational distribution and calculate $\log\ p(y; \Theta)$ as in (38) T. Li and Ma (2023).

$$\begin{aligned} log\, p(y; \Theta) &= \int q(x)\ \log\,(\frac{p(y, x; \Theta)}{q(x)})\, dx - \int q(x)\ \log\,(\frac{p(x|y; \Theta)}{q(x)})\, dx \\ &= \mathcal{L}(q(x); \Theta) + \mathcal{KL}(q(x)\,||\,p(x|y; \Theta)) \end{aligned} \tag{38}$$

By moving the $\mathcal{D}_{\mathcal{KL}}(.\,||\,.)$ to the left side of the equation we would have:

$$\mathcal{L}(q(x); \Theta) = \log\ p(y; \Theta) - \mathcal{KL}(q(x)\,||\,p(x|y; \Theta)) \tag{39}$$

Since the Kullback-Leibler divergence between the two distributions is non-negative, $\mathcal{L}(q(x); \Theta)$ would be the lower bound (ELBO) on the log-likelihood. Therefore, instead of maximizing the log-likelihood, we can maximize the ELBO T. Li and Ma (2023); Verbeek, Vlassis, and Nunnink (2003). Using Jensen's inequality the lower bound would be derived as follows:

$$\mathcal{L}(q(x);\Theta) = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{E}_q[\log\, p(w_{d,n}|z_{d,n},\beta)\, p(z_{d,n}|\theta,c_d)] + \mathbb{E}_q[\log\, p(\pi|\alpha)] \\ + \sum_{m=1}^{M}\mathbb{E}_q[\log\, p(\theta_m|\lambda)] + \sum_{d=1}^{D}\mathbb{E}_q[\log\, p(c_d|\pi)] - \mathbb{E}_q[\log\, q(\pi,\theta,c,z)] \tag{40}$$

Consequently, the E-step and the M-step are calculated by setting the partial derivatives concerning each variational and model parameter to zero. Since the goal is to propose a probabilistic projection-clustering model, these equations are divided into two clustering and projection parts.

Clustering task

The clustering task is related to the computation of the latent variables ($\psi_{d,m}$, $\gamma_{m,k}$, and η_m) corresponding to the clustering variables which are cluster centers θ_m, document indicators c_d, and the mixing weights π. Following, the updating equations for clustering and the definition for each latent variable are represented.

$$\psi_{d,m} = \exp\left\{\sum_{k=1}^{K}\left[\Big(\Psi(\gamma_{m,k}) - \Psi(\sum_{i=1}^{K}\gamma_{m,i})\Big)\sum_{n=1}^{N_d}\phi_{d,n,k}\right] + \Psi(\eta_m) - \Psi(\sum_{i=1}^{M}\eta_i)\right\} \tag{41}$$

$$\gamma_{m,k} = \sum_{d=1}^{D}\psi_{d,m}\sum_{n=1}^{N_d}\phi_{d,n,k} + \lambda_k \tag{42}$$

$$\eta_m = \sum_{d=1}^{D}\psi_{d,m} + \frac{\alpha}{M} \tag{43}$$

$\psi_{d,m}$ is a posteriori probability referring to the probability that document *d* belongs to the cluster *m* among *M* clusters ($p(c_d = m)$). Cluster centers are represented by $\gamma_{m,k}$ and it can be considered as the *k*th component of θ_m in the topic space. Note that here, $\vec{\phi}$ is considered to be a constant vector as it related to the words projection and represents the a posteriori probability that word $w_{d,n}$ is sampled from topic *k* in the topic space. Lastly, η_m controls the mixing weights π for clusters and determines the probability of cluster m Yu et al. (2005).

Projection task

In the projection phase, we consider $\vec{\psi}$, $\vec{\gamma}$, and $\vec{\eta}$ to be fixed, thus the variational parameter $\vec{\phi}$ that is referring to a posteriori probability that word $w_{d,n}$ is sampled from topic k, and the projection parameter $\vec{\beta}$ would be updated.

$$\phi_{d,n,k} = \beta_{k,w_{d,n}} \exp\left\{ \sum_{m=1}^{M} \psi_{d,m} \left[\Psi(\gamma_{m,k}) - \Psi(\sum_{i=1}^{K} \gamma_{m,i}) \right] \right\} \tag{44}$$

$$\beta_{k,j} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,k}\, \delta_j(w_{d,n}) \tag{45}$$

$\delta_j(w_{d,n})$ is a binary variable, it equals 1 if the index j is associated with the word $w_{d,n}$, otherwise, it is 0 Yu et al. (2005).

In both clustering and projection parts, we iterate through the equations until convergence to reach the maximum data likelihood as it is in the variational EM algorithm.

A Dirichlet prior $Dir(\lambda)$ was proposed for all the cluster centers $\vec{\theta}$ in this model. The Dirichlet distribution is commonly used as a prior for the multinomial distribution, known for its flexibility and simplicity Bouguila (2009); Koochemeshkian, Zamzami, and Bouguila (2020), but it has some limitations. The negative covariance structure and the constraint of the constant sum are some of the most significant limitations of the Dirichlet distribution which can be restrictive in important applications Zamzami and Bouguila (2019b). In this study, we propose using the generalized Dirichlet distribution and the Beta-Liouville distribution as alternative priors to examine the complexity of the model. In the next section, we would review the generalized Dirichlet distribution and the Beta-Liouville distribution and update the equations regarding those distributions.

3.3 Proposed Approaches

In this section, we briefly review the generalized Dirichlet distribution and the Beta-Liouville distribution since we are using them as the alternative prior distributions, and update the equations based on the new priors.

3.3.1 Probabilistic clustering-projection model using GD prior

To overcome the restrictions of the Dirichlet distribution Bouguila and ElGuebaly (2008a, 2008b); Bouguila and Ghimire (2010), Connor and Mosimann Connor and Mosimann (1969) introduced the generalized Dirichlet distribution by using the concept of neutrality. A random vector $\vec{X}$ is considered to be completely neutral, if for all values of j ($j < K$), $(x_1, x_2, ..., x_j)$ is independent of $(x_{j+1}, x_{j+2}, ..., x_K)/(1 - \sum(x_1, x_2, ..., x_j))$. By supposing a beta distribution with parameters α_{j+1} and β_{j+1} for each element of $(x_{j+1}, x_{j+2}, ..., x_K)/(1 - \sum(x_1, x_2, ..., x_j))$, the density function for the generalized Dirichlet distribution was derived as follows:

$$GD(\vec{X}|\lambda_1, ..., \lambda_K, \rho_1, ..., \rho_K) = \prod_{i=1}^{K} \frac{\Gamma(\lambda_i + \rho_i)}{\Gamma(\lambda_i)\Gamma(\rho_i)} x_i^{(\lambda_i - 1)} (1 - \sum_{j=1}^{i} x_j)^{\gamma_i} \tag{46}$$

where for $i = 1, 2, ..., K$, $x_i \geq 0$ and $\sum_{i=1}^{K} x_i \leq 1$, and for $i = 1, 2, ..., K-1$, $\gamma_i = \rho_i - (\lambda_{i+1} + \rho_{i+1})$, and $\gamma_K = \rho_K - 1$ Epaillard and Bouguila (2019); Wong (2009).

When $\beta_i = \lambda_{i+1} + \rho_{i+1}$, the generalized Dirichlet distribution is reduced to the Dirichlet distribution. Therefore, we can conclude that the Dirichlet distribution is a special case of the generalized Dirichlet distribution Bouguila (2008); Bouguila and Ziou (2004b). Similar to the Dirichlet distribution, the GD distribution is also a conjugate to the multinomial distribution Bouguila and Ghimire (2010); Najar and Bouguila (2022a). However, despite the Dirichlet distribution, the GD distribution has a more general covariance structure; meaning that variables with the same means do not necessarily have the same covariance, and unlike the Dirichlet distribution covariance between two variables is not negative Najar and Bouguila (2021a). These properties of the GD distribution make it a good prior for Bayesian classifiers Koochemeshkian et al. (2020).

For this study, we put a GD prior GD($\vec{\lambda}$,$\vec{\rho}$) for all the cluster centres, where $\vec{\lambda} = [\lambda_1, ..., \lambda_K]$ and $\vec{\rho} = [\rho_1, ..., \rho_K]$ and we show it as θ_{GD}. Therefore, equation (40) would be updated as follows:

$$\begin{aligned}\mathcal{L}(q(x); \Theta) = &\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log\, p(w_{d,n}|z_{d,n}, \beta)\, p(z_{d,n}|\theta_{GD}, c_d)] + \sum_{m=1}^{M} \mathbb{E}_q[\log\, p(\theta_{GD(m)}|\vec{\lambda}, \vec{\rho})] \\ &+ \mathbb{E}_q[\log\, p(\pi|\alpha)] + \sum_{d=1}^{D} \mathbb{E}_q[\log\, p(c_d|\pi)] - \mathbb{E}_q[\log\, q(\pi, \theta_{GD}, c, z)]\end{aligned} \tag{47}$$

where $p(\theta_{GD(m)}|\vec{\lambda},\vec{\rho})$ represents the probability density function of a K-dimensional vector $\vec{\theta_m} = [X_{m_1}, X_{m_2}, \ldots, X_{m_K}]$ following a GD distribution with positive parameter vectors $\vec{\lambda}$ and $\vec{\rho}$.

It can be proved that the GD distribution is part of the exponential family Bouguila (2011a); Zamzami and Bouguila (2019a, 2019b, 2022), so its expected value would be calculated by taking the derivative of its cumulant function $A(\eta)$ as it is shown in Appendix A.1.

3.3.2 Probabilistic clustering-projection model using BL prior

In addition to the GD distribution, in this study we will utilize a member of the Liouville distributions called Beta-Liouville distribution as a prior for the cluster centres θ_m. Studies showed that the BL distribution is a good alternative compared to the Dirichlet and GD distributions in the statistical presentation of proportional data. The BL distribution is also a conjugate to the multinomial distribution, and similar to the GD distribution has a more general covariance structure than the Dirichlet distribution Zamzami and Bouguila (2019b). The probability density function of a random vector $\vec{X}$ following a BL distribution with positive parameters $(\lambda_1, \lambda_2, ..., \lambda_K, \lambda, \nu)$ is:

$$\begin{aligned} BL(\vec{X}|\lambda_1, \lambda_2, ..., \lambda_K, \lambda, \nu) = \Gamma(\sum_{k=1}^{K}\lambda_k)\frac{\Gamma(\lambda+\nu)}{\Gamma(\lambda)\Gamma(\nu)}\prod_{k=1}^{K}\frac{X_k^{(\lambda_k-1)}}{\Gamma(\lambda_k)} \\ \times \left(\sum_{k=1}^{K}X_k\right)^{\lambda-\sum_{k=1}^{K}\lambda_k}\left(1-\sum_{k=1}^{K}X_k\right)^{\nu-1} \end{aligned} \tag{48}$$

Note that the Beta-Liouville includes Dirichlet as a special case Bouguila (2010); Fan and Bouguila (2015). Interested readers are referred to Bakhtiari and Bouguila (2012); Bouguila (2010); Daghyani et al. (2019); Fan and Bouguila (2013) for more discussions about Beta-Liouville distribution. Like the previous part, we put a BL prior for all the cluster centres $\theta_m \sim BL(\lambda_1, \lambda_2, ;\lambda_K, \lambda, \nu)$ and update (40) accordingly:

$$\begin{aligned} \mathcal{L}(q(x);\Theta) = &\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{E}_q[\log\, p(w_{d,n}|z_{d,n},\beta)\, p(z_{d,n}|\theta_{BL},c_d)] \\ &+\sum_{m=1}^{M}\mathbb{E}_q[\log\, p(\theta_{BL(m)}|\lambda_1,\lambda_2,;\lambda_K,\lambda,\nu)] + \mathbb{E}_q[\log\, p(\pi|\alpha)] \\ &+\sum_{d=1}^{D}\mathbb{E}_q[\log\, p(c_d|\pi)] - \mathbb{E}_q[\log\, q(\pi,\theta_{BL},c,z)] \end{aligned} \tag{49}$$

$p(\theta_{BL(m)}|\lambda_1, \lambda_2, ;\lambda_K, \lambda, \nu)$ denotes the probability density function of a K-dimensional vector θ_m ($\vec{\theta_m} = [X_{m_1}, X_{m_2}, \ldots, X_{m_K}]$) following BL distribution with parameters $[\lambda_1, \lambda_2, ..., \lambda_K, \lambda, \nu]$.

Subsequently, since the BL distribution is part of the exponential family Bouguila (2011a); Zamzami and Bouguila (2020b), therefore same as the GD, the expected value of the BL distribution is calculated by taking the derivative of its cumulant function $A(\eta)$ (Appendix A.2).

3.4 Experimental Results

The purpose of this study is to determine the performance of the GD prior and BL prior in a probabilistic clustering projection algorithm. In the previous study, the traditional projection and clustering methods were compared with the PCP model using the Dirichlet distribution as a prior for the cluster centers. The results showed that in the case of projection, the PCP model has less complexity rather than the traditional PLSI and LDA models. On the other hand, in document clustering, PCP outperformed the NMF algorithm and K-means. However, choosing Dirichlet as a prior comes with limitations. As a result, GD and BL distributions have been suggested as alternative priors to overcome such restrictions.

This section presents experimental outcomes for the PCP model, which are compared with other models based on three criteria: document modeling, word projection, and document clustering. Two text data collections are used for comparison: Reuters-21578, in which we selected 5 categories earn, acq, money-fx, grain, crude, and a four-group sample from 20Newsgroup consisting of autos, motorcycles, baseball, and hockey. The first step for both data collections is preprocessing including removing stop words and the words that occurred in less than 5 documents, tokenization, and stemming. As a result, 8210 documents and 4856 unique words are obtained for the Reuters-21578 dataset and 3979 documents and 6175 words are acquired from 20Newsgroup.

3.4.1 Document Modelling

In this study, we compare the performance of the PCP model using three different priors: Dirichlet, generalized Dirichlet, and Beta-Liouville. As in the previous study Yu et al. (2005), we use perplexity as the comparison metric and split the dataset into 80% training data and 20% testing

data. We calculate perplexity for each model using the same method as in the previous study (Equation 50), with the exception that we now use the respective priors for each model. We optimize the smoothing term for each model and train them until the improvement is less than 0.01%. Tables 3.1 and 3.2 show the comparison of the perplexity between our different models for the different number of topics (K). It is illustrated that the PCP model using the GD prior outperforms the other two PCP models, indicating a better fit to the data.

$$Prep(\mathcal{D}) = \exp\left(-\log p(\mathcal{D}) / \sum_{d} |w_d|\right) \tag{50}$$

where $|w_d|$, refers to the length of the document d. The lower the perplexity is, the better fit the model would be.

Table 3.1: Calculated Perplexity for Reuters-21578 Dataset

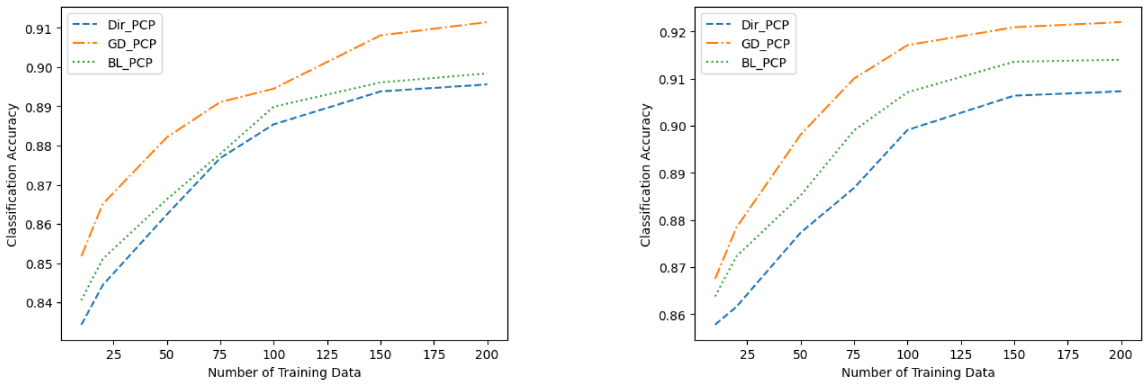
Reuters-21578					
K	10	20	30	40	50
Dir_PCP	1046.61	1009.84	965.63	772.85	666.68
GD_PCP	888.74	796.74	658.4	543.67	489.67
BL_PCP	987.33	866.59	832.33	768.97	536.25

Table 3.2: Calculated Perplexity for 20Newsgroup Dataset

20Newsgroup					
K	10	20	30	40	50
Dir_PCP	1992.88	1883.98	1692.91	1568.11	1455.45
GD_PCP	1692.9	1663.23	1551.89	1395.09	1265.8
BL_PCP	1894.18	1727.75	1564.51	1474.46	1306.5

3.4.2 Word Projection

As mentioned in Section 2 of this chapter, we can explain PCP as a projection model when the clustering structure is known. To evaluate the quality of the models, we used the same method as in the previous study, we split the data into training and testing sets, set the number of topics to 10, and for each category, we create an SVM classifier on the low-dimensional representations of



(a) Reuters

(b) 20Newsgroup

Figure 3.1: Accuracy of different models in Reuters (a) and 20Newsgroup (b) datasets.

the models, which for the PCP model is the projection term $\sum_{n=1}^{N_d} \phi_{d,n,k}$, and fit the classifier to the training data. We then use the trained classifier to predict the class labels for the test data and calculate the accuracy of the classifier for each category. Finally, we calculated the average accuracy over all categories. We repeated the steps for different numbers of training data as it is shown on the x-axis in Figure 3.1. It is shown that employing generalized Dirichlet prior provides a better word projection compared to the Beta-Liouville prior and Dirichlet prior.

3.4.3 Document Clustering

In this section, we are studying the ability of the PCP model on clustering the documents. For that purpose, we use the normalized mutual information $\widehat{MI}$ as an evaluation metric to measure the similarity between the generated clustered data and the actual clusters Ji and Xu (2006); Xu, Liu, and Gong (2003).

Consider X and Y as two sets of document clusters, the mutual information (MI) measures the relationship between these clusters using a joint probability mass function $p(x, y)$, along with the marginal probabilities $p(x)$ and $p(y)$ and is defined as follows:

$$MI(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{51}$$

The mutual information ranges from zero to a maximum value equal to the higher entropy between $H(X)$ and $H(Y)$. Entropy represents the amount of uncertainty or randomness in X and

Y. When the two sets of document clusters, X and Y, are identical, the mutual information reaches its maximum value, which is equal to the higher entropy of either X or Y. On the other hand, when the two sets are completely independent, the mutual information becomes zero, indicating no shared information between them.

Instead of using mutual information ($MI(X,Y)$), we use a normalized measure termed $\widehat{MI}$ (equation 52) to make it easier to compare various pairs of cluster sets. With a range of zero to one, this measure makes it simpler to read and analyze the data Ji and Xu (2006).

$$\widehat{MI}(X,Y) = \frac{MI(X,Y)}{\max(H(X),H(Y))} \tag{52}$$

The $\widehat{MI}$ is calculated for different methods we employed in this study, as illustrated in Table 3.3, using the true cluster numbers, which is 5 for Reuters data collection, and 4 for Newsgroup.

Table 3.3: Document Clustering Comparison

	Dir_PCP	GD_PCP	BL_PCP
$Reuters$	0.406	0.578	0.471
$20Newsgroup$	0.58	0.721	0.648

3.5 Conclusion

In this chapter, we proposed a probabilistic clustering-projection model using three different distributions as a prior, namely Dirichlet, generalized Dirichlet, and Beta-Liouville, and aimed to investigate the combined influence of clustering and projection within a single model, while also exploring the impact of different prior knowledge on information retrieval. The experiments were conducted on two text datasets, Reuters-21578 and 20Newsgroup, and analyzed document modeling, word projection, and document clustering. Results showed improvements in the case of generalized Dirichlet and Beta-Liouville as a prior rather than Dirichlet distribution. This research provides valuable insights into enhancing the accuracy and effectiveness of information retrieval methods, highlighting the potential benefits of incorporating diverse prior knowledge in similar studies.

Chapter 4

Generalized Conditional Naive Bayes model

4.1 Introduction

In the realm of unsupervised learning, the structure of the data remains hidden from the observer which prompted the development of probabilistic mixture models. Indeed, a powerful approach aimed at figuring out this hidden structure, given that the data comprises a mixture of multiple underlying components X. Li, Ling, and Wang (2016). Naive-Bayes (NB) models are a type of generative mixture models known for their simplicity, accuracy, and speed, making them widely used in tasks like product recommendations, medical diagnoses, software defect predictions, and cybersecurity. In addition, researchers have shown that these models tend to outperform other approaches, such as C4.5, PEBLS, and CN2 classifiers, especially in cases with small datasets Domingos and Pazzani (1997); Wickramasinghe and Kalutarage (2021). In the era of big data, it is common to encounter issues like sparsity, missing values, and unobserved data. This is often due to the fact that users have limited knowledge about the vast number of available items. Hence, employing traditional NB models won't be advantageous when dealing with such large-scale datasets. To tackle the sparsity problem, a generalized form of the Naive Bayes model, referred to as the conditional Naive Bayes (CNB) model, was introduced Taheri, Mammadov, and Bagirov (2010). This model

calculates the likelihood of each class for a given feature vector by utilizing a subset of observed features, rather than incorporating all of them, thus addressing the sparsity problem. However, unlike the traditional Naive Bayes model, the CNB model does not consider the assumption of feature independence. To tackle this limitation, alternative models were suggested including multi-case model Sahami, Hearst, and Saund (1996), overlapping mixture model Fu and Banerjee (2008); Kyoya and Yamanishi (2021), aspect model Hofmann (2001); Minka and Lafferty (2012), LDA D. M. Blei et al. (2003), and LD-CNB model Banerjee and Shan (2007). Latent Dirichlet Allocation (LDA) is a probabilistic generative model of a corpus, where documents are represented as random mixtures over latent low-dimensional topic space. Assuming K latent topics, a document is generated by sampling a mixture of these topics, with each topic represented as a probability distribution over the words in the document, and then sampling words from that mixture. The key aspect of LDA is that despite the CNB model, it allows documents to be associated with two or more topics D. Blei, Ng, and Jordan (2001); D. M. Blei et al. (2003). The latent Dirichlet conditional Naive-Bayes (LD-CNB) model was presented as a more adaptable model since it utilizes exponential family distribution in variational approximation for model inference and learning. In the research conducted by Banerjee et al. Banerjee and Shan (2007), they applied Gaussian and Discrete distributions as specific examples of such exponential family distributions. Through a comparison between the LD-CNB and the CNB models, it has been demonstrated that the LD-CNB model consistently outperforms the CNB model in terms of having lower perplexity. However, using Dirichlet as a prior distribution in the model can lead to some constraints. To address the limitations associated with the constricting negative covariance structure of Dirichlet distribution, this paper introduces an approach where we suggest employing alternative distributions, specifically the generalized Dirichlet (GD) distribution and the Beta-Liouville (BL) distribution, as priors to define the mixing weights for the data point in the model.

The chapters's organization is as follows: In section 2, we provide an overview of the LD-CNB model, its instantiations for exponential family distributions such as Gaussian and Discrete distributions, and the variational Expectation Maximization algorithm used for learning and inference. Section 3 covers a review of the properties of the generalized Dirichlet (GD) distribution and the Beta-Liouville (BL) distribution, our proposed approaches, and the updated model based on each of

those prior distributions. Section 4 presents the experimental results obtained from the UCI benchmark repository Frank (2010) and Movielens recommendation system datasetHarper and Konstan (2015). Finally, in section 5, we offer our conclusions.

4.2 Latent Dirichlet Conditional Naive Bayes

In this section, we will examine the LD-CNB model and discuss the constraints of both the LDA model and the Naive-Bayes model, which led to the development of LD-CNB. Additionally, we will delve into the details of the variational EM algorithm and the computational steps taken to accomplish the goals of model learning and inference.

The LD-CNB model was proposed in response to the limitations of NB models in handling sparsity within large-scale data sets. Because the observer has limited knowledge regarding the magnitude of the items, the likelihood of encountering missing or unobserved values rises. Although NB models demonstrated their accuracy and ability in processing small datasets, they are still not able to handle the sparsity in the case of big data. Furthermore, in the NB model, it is assumed that features come from a single mixture component, which imposes significant limitations on the modeling capabilities of the NB model.

In order to address the challenges associated with sparsity, the Conditional Naive-Bayes (CNB) model was introduced. This model conditions a Naive-Bayes model on only a subset of observed features. Let's assume that d represents the total number of features in the dataset, a subset of features is denoted as $f = \{f_1, ..., f_m\}$, where $m < d$. The conditional probability of the feature vector x is then computed as follows:

$$p(x|\pi, \Theta, f) = \sum_{z=1}^{K} p(z|\pi) \prod_{j=1}^{m} p_\psi(x_j|z, \Theta, f_j) \tag{53}$$

where π represents prior distribution over K components. The term ψ refers to the appropriate exponential family model for feature f_j and $p_\psi(x_j|z, \Theta, f_j)$ is the exponential family distribution for f_j. $z = (1, ..., K)$ and $\Theta = \{\theta_z\}$ are defined as the parameters for the exponential family distribution.

In the context of LDA, a 'data point' is presented as a sequence of tokens (feature), with each

token generated from the same discrete distribution, since they are considered semantically identical Griffiths and Steyvers (2004). In some applications, instead of considering a feature as a token, each feature is associated with a measured value, which can be real or categorical. Besides that, various features within the feature set can carry distinct semantic meanings. Because the NB model assumes that features come from the same mixture component, they took a Dirichlet prior with parameter α for the mixing weight π to overcome the problem caused by that assumption. Therefore, the process of generating a sample x following the LD-CNB model can be outlined as follows:

(1) Choose $\pi \sim \text{Dir}(\alpha)$

(2) For each of the observed feature f_j (j=1, ..., m):

(a) Choose $z_j \sim \text{Discrete}(\pi)$

(b) Choose a feature value $x_j \sim p_\psi(x_j|z_j, \Theta, f_j)$

When taking the model parameters into account, the joint distribution of (π, z, x) can be expressed as:

$$p(\pi, z, x|\alpha, \Theta, f) = p(\pi|\alpha) \prod_{j=1}^{m} p(z_j|\pi) p_\psi(x_j|z_j, \Theta, f_j) \tag{54}$$

Given the feature set of the entire data set denoted as $F = \{f_1, ..., f_N\}$, the probability of the entire data set $X = \{x_1, ..., x_N\}$ can be calculated as follows:

$$p(X|\alpha, \Theta, f) = \prod_{i=1}^{N} \int_\pi p(\pi|\alpha) \left(\prod_{j=1}^{m_i} \sum_{z_{ij}=1}^{K} p(z_{ij}|\pi) p_\psi(x_{ij}|z_{ij}, \Theta, f_{ij}) \right) d\pi \tag{55}$$

It can be seen from equation 55, that the model is dependent on the observed features and their potential values. Thus, when generating the value x_j for the feature f_j, it is necessary to select the suitable exponential family model (ψ). It's important to note that the choice of family distribution depends on the specific feature because each feature may have a different family distribution.

In the research conducted by Banerjee et al. Banerjee and Shan (2007), they utilized a univariate Gaussian distribution for real-valued features and a Discrete distribution for categorical features within each class. For the Gaussian distribution model (LD-CNB-Gaussian), the model parameters

are denoted as $\Theta = \{(\mu_{(z,f_j)}, \sigma^2_{(z,f_j)})\}$, where $j = 1, ..., d$, and $z = 1, ..., K$ (d and K representing the total number of features and the number of latent classes in the dataset, respectively). Therefore, in equation 55, $p_\psi(x_{ij}|z_{ij}, \Theta, f_{ij})$ can be updated as $p(x_j|\mu_{(z,f_j)}, \sigma^2_{(z,f_j)})$. In the case of Discrete distribution (LD-CNB-Discrete model), each feature is allowed to be of a different type and a different number of possible values. Assuming K latent classes ($z = 1, ..., K$), and d features with r_j ($j = 1, ..., d$) possible values for each feature, the model parameters for latent class z and feature f_j are represented by discrete probability distribution over possible values $\Theta = \{p_{(z,f_j)}(r)\}$, where $r = (1, ..., r_j)$.

4.2.1 Model Learning and Inference

Variational EM Algorithm

Consider y as the observed data generated through a set of latent variables x. Let Θ denotes the model parameter describing the dependencies between variables. Consequently, the likelihood of observing the data can be expressed as a function of Θ. The objective is to identify the optimal value for Θ that maximizes the likelihood, or equivalently, the logarithm of the likelihood, as illustrated in equation 56.

$$\log\, p(y|\Theta) = \log \int p(x, y|\Theta)\, dx \tag{56}$$

However, the computation of maximum log-likelihood is typically a complex task. As a solution, an arbitrary distribution for hidden variables, denoted as $q(x)$, is defined. The marginal likelihood can then be broken down with respect to $q(x)$ as outlined below:

$$\begin{aligned} \log\, p(y|\Theta) &= \log \int q(x)\, \frac{p(x, y|\Theta)}{q(x)}\, dx - \log \int q(x)\, \frac{p(x|y, \Theta)}{q(x)}\, dx \\ &= \mathcal{L}(q(x)|\Theta) + \mathcal{KL}(q(x)||p(x|y, \Theta)) \end{aligned} \tag{57}$$

The term $\mathcal{L}(q(x)|\Theta)$ is referred to as the evidence lower bound (ELBO), serving as a lower bound for $\log p(y|\Theta)$ due to the non-negativity of $\mathcal{KL}(q(x)||p(x|y, \Theta))$ T. Li and Ma (2023). To achieve the maximum log-likelihood, we can either minimize $\mathcal{KL}(q(x)||p(x|y, \Theta))$ or maximize

the evidence lower bound (ELBO), denoted as $\mathcal{L}(q(x)|\Theta)$. Consequently, rather than directly maximizing the log-likelihood, the focus is on maximizing the ELBO T. Li and Ma (2023); Verbeek et al. (2003). This approach leads to the development of a variational EM algorithm, which iteratively optimizes the lower bound of the log-likelihood.

$$q(\pi, z|\gamma, \phi, f) = q(\pi|\gamma) \prod_{j=1}^{m} q(z_j|\phi_j) \tag{58}$$

$q(\pi, z|\gamma, \phi, f)$ is introduced as a variational distribution over the latent variables conditioned on free parameters γ and ϕ, where γ is a Dirichlet parameter, and $\phi = (\phi_1, ..., \phi_m)$ is a vector of multinomial parameters.

Based on the information above, the associated ELBO can be computed as follows:

$$\mathcal{L}(\gamma, \phi; \alpha, \Theta) = \mathbb{E}_q[\log\ p(\pi|\alpha)] + \mathbb{E}_q[\log\ p(z|\pi)] + \mathbb{E}_q[\log\ p(x|z, \Theta)] + \mathcal{H}(q(\pi)) + \mathcal{H}(q(z)) \tag{59}$$

The variational EM-step is derived by setting the partial derivatives, with respect to each variational and model parameter, to zero. The ELBO can be optimized iteratively by employing the following set of update equations:

$$\phi_{(z_j, f_j)} \propto \exp\left(\Psi(\gamma_{z_j}) - \Psi(\sum_{z_{j'}=1}^{K} \gamma_{z_{j'}})\right) p_\psi(x_j|z_j, \Theta, f_j) \tag{60}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{61}$$

As previously shown, the respective distributions for LD-CNB-Gaussian and LD-CNB-Discrete will be substituted with $p_\psi(x_j|z_j, \Theta, f_j)$ in (60). The updated corresponding parameter (Θ) for each model is then calculated as follows:

- LD-CNB-Gaussain

$$\mu_{(z_j, f_j)} = \frac{\sum_{i=1}^{N} \phi_{i(z_j, f_j)} x_{ij}}{\sum_{i=1}^{N} \phi_{i(z_j, f_j)}} \tag{62}$$

$$\sigma^2_{(z_j,f_j)} = \frac{\sum_{i=1}^{N} \phi_{i(z_j,f_j)}(x_{ij} - \mu_{(z_j,f_j)})^2}{\sum_{i=1}^{N} \phi_{i(z_j,f_j)}} \tag{63}$$

- LD-CNB-Discrete

$$p_{(z_j,f_j)}(r) \propto \sum_{i=1}^{N} \phi_{i(z_j,f_j)} x_{ij} \mathbb{1}(r|i,f_j) + \epsilon \tag{64}$$

In equation 64, the term $\mathbb{1}(r|i,f_j)$ refers to the indicator matrix of observed value r for feature f_j in observation x_i.

However, using Dirichlet as a prior presents some restrictions, especially when modeling correlated topics. First, all data features are bound to share a common variance, and their sum must be equal to one. Consequently, we cannot introduce individual variance information for each component of the random vector. In addition, when using a Dirichlet distribution, we have only one degree of freedom to convey our confidence in the prior knowledge. All the entries in the Dirichlet prior are always negatively correlated which means if the probability of one component increases, the probabilities of the other components must either decrease or remain the same to ensure they still sum up to one Caballero et al. (2012); Eita, Ennajari, and Bouguila (2023). These limitations motivated us to employ a generalized form of Dirichlet distribution, namely generalized Dirichlet distribution, and Beta-Liouville distribution as potential priors for the multinomial distribution.

4.3 Proposed Approaches

In this section, we provide a concise overview of the generalized Dirichlet distribution and the Beta-Liouville distribution. We then proceed to adjust the equations in accordance with these new priors.

4.3.1 Latent Generalized Dirichlet Conditional Naive Bayes

To overcome the limitations associated with the Dirichlet distribution, Bouguila and ElGuebaly (2008a, 2008b); Bouguila and Ghimire (2010), Connor and Mosimann introduced the concept of neutrality and developed the generalized Dirichlet distribution Connor and Mosimann (1969)

is conjugate to the multinomial distribution Bouguila and Ghimire (2010); Najar and Bouguila (2022a). In this context, a random vector $\vec{X}$ is considered completely neutral when, for all values of j ($j < K$), the vector $(x_1, x_2, ..., x_j)$ is independent of the vector $(x_{j+1}, x_{j+2}, ..., x_K)/(1 - \sum_j (x_1, x_2, ..., x_j))$, which means that a neutral vector does not impact the proportional division of the remaining interval among the rest of the variables. By assuming a univariate beta distribution with parameters α and β for each component of $(x_1, x_2, ..., x_{K-1})$, the probability density function for the generalized Dirichlet distribution is derived as follows:

$$GD(\overrightarrow{X}|\alpha_1, ..., \alpha_K, \beta_1, ..., \beta_K) = \prod_{i=1}^{K} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{(\alpha_i - 1)} (1 - \sum_{j=1}^{i} x_j)^{\gamma_i} \tag{65}$$

where

$$\text{for } \; i = 1, 2, ..., K-1, \;\; \gamma_i = \beta_i - (\alpha_{i+1} + \beta_{i+1})$$

$$\text{and } \; \gamma_K = \beta_K - 1.$$

Note that $\alpha_i, \beta_i > 0$. For $i = 1, 2, ..., K$, $x_i \geq 0$ and $\sum_{i=1}^{K} x_i \leq 1$ Epaillard and Bouguila (2019); Wong (2009). Assuming $\beta_{i-1} = \alpha_i + \beta_i$ the generalized Dirichlet distribution is reduced to the Dirichlet distribution, which indicates Dirichlet distribution as a special case of the generalized Dirichlet distribution Bouguila (2008); Bouguila and Ziou (2004b). The mean, the variance, and the covariance in the case of the generalized Dirichlet distribution, for $i = 1, ..., K-1$ are as follows:

$$E(X_i) = \frac{\alpha_i}{\alpha_i + \beta_i} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j} \tag{66}$$

$$Var(X_i) = E(X_i) \left(\frac{\alpha_i + 1}{\alpha_i + \beta_i + 1} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1} - E(X_i) \right) \tag{67}$$

$$COV(X_i, X_d) = E(X_d) \left(\frac{\alpha_i}{\alpha_i + \beta_i + 1} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1} - E(X_i) \right) \tag{68}$$

Unlike Dirichlet distribution, the GD distribution has a more general covariance structure, and variables with the same means are not obligated to have the same covariance. Moreover, for GD distribution covariance between two variables is not negative Najar and Bouguila (2021a). This

flexibility and properties of the GD distribution make it desirable prior to the topic modeling and finding the hidden structure of the data Koochemeshkian et al. (2020).

4.3.2 Model Learning and Inference

Variational EM algorithm

In the proposed approach, we consider a GD prior with parameters α, β for the mixing weights of the data points of the model ($\pi \sim GD(\alpha, \beta)$), and Θ as the model parameter. Therefore, the joint distribution of (π, z, x) is calculated as:

$$p(\pi, z, x|\alpha, \beta, \Theta, f) = p(\pi|\alpha, \beta) \prod_{j=1}^{m} p(z_j|\pi) p_\psi(x_j|z_j, \Theta, f_j) \tag{69}$$

Following that, the variational distribution for updated model parameters is defined as:

$$q(\pi, z|\gamma, \lambda, \phi, f) = q(\pi|\gamma, \lambda) \prod_{j=1}^{m} q(z_j|\phi_j) \tag{70}$$

where γ and λ are the parameters for the generalized Dirichlet distribution, and $\phi = (\phi_1, ..., \phi_m)$ denotes a vector of parameters for the multinomial distribution.

Further, in order to determine the maximum likelihood of the data, we seek to maximize the associated lower bound (ELBO), computed as follows:

$$\begin{aligned} \mathcal{L}(\gamma, \lambda, \phi; \alpha, \beta, \Theta) = \mathbb{E}_q[\log\, p(\pi|\alpha, \beta)] + \mathbb{E}_q[\log\, p(z|\pi)] + \mathbb{E}_q[\log\, p(x|z, \Theta)] \\ + \mathcal{H}(q(\pi)) + \mathcal{H}(q(z)) \end{aligned} \tag{71}$$

It has been demonstrated that the GD distribution belongs to the exponential family, so its expected value is calculated by taking the derivative of its cumulant function (Appendix A.1). By setting the partial derivatives to zero with regard to each parameter and subsequently deriving the revised equations for variational and model parameters (equations 72,73, and 74), we can find the maximum value for the ELBO.

$$\phi_{(z_j, f_j)} \propto p_\psi(x_j|z_j, \Theta, f_j) \exp\left(\Psi(\gamma_{z_j}) - \Psi(\lambda_{z_j}) - (\sum_{i=1}^{z_j} \Psi(\gamma_i + \lambda_i) - \Psi(\lambda_i))\right) \tag{72}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{73}$$

$$\lambda_{z_j} = \beta_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{74}$$

Given that the model parameter Θ form is independent of the prior distribution, the update equations for exponential family parameters remain unchanged, as presented in (62,63, and 64).

4.3.3 Latent Beta-Liouville Conditional Naive Bayes

In this section, we will incorporate another distribution known as the Beta-Liouville (BL) distribution as a prior in our model. Research has demonstrated that the BL distribution offers a viable alternative to the Dirichlet and GD distributions for statistically representing proportional data. The BL distribution also serves as a conjugate prior for the multinomial distribution and, similar to the GD distribution, it has a more general covariance structure Luo, Amayri, Fan, and Bouguila (2023). The probability density function for a random vector $\vec{X}$ following a BL distribution with positive parameter vector $\vec{\Phi} = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ is expressed as:

$$BL(\vec{X}|\vec{\Phi}) = \Gamma(\sum_{k=1}^{K} \alpha_k) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{k=1}^{K} \frac{X_k^{(\alpha_k-1)}}{\Gamma(\alpha_k)} \left(\sum_{k=1}^{K} X_k\right)^{\alpha-\sum_{k=1}^{K}\alpha_k} \left(1 - \sum_{k=1}^{K} X_k\right)^{\beta-1} \tag{75}$$

The Beta-Liouville distribution transforms into the Dirichlet distribution when the generator density follows a beta distribution with parameters $\sum_{i=1}^{K-1} \alpha_i$ and α_K, as explained in Bouguila (2010); Fan and Bouguila (2015). The mean, the variance, and the covariance of the Beta-Liouville distribution are calculated as follows:

$$E(X_i) = \frac{\alpha}{\alpha+\beta} \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{76}$$

$$Var(X_i) = E(X_i) \left(\frac{\alpha+1}{\alpha+\beta+1} \frac{\alpha_k+1}{\sum_{k=1}^{K}\alpha_k+1}\right) - E(X_i)^2 \left(\frac{\alpha_k^2}{(\sum_{k=1}^{K}\alpha_k)^2}\right) \tag{77}$$

$$COV(X_i, X_d) = \frac{\alpha_i \alpha_d}{\sum_{i=1}^{K}\alpha_i} \left(\frac{-\alpha^2}{(\alpha+\beta)^2 \sum_{k=1}^{K}\alpha_k} + \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)(\sum_{k=1}^{K}\alpha_k+1)}\right) \tag{78}$$

By considering a Beta-Liouville (BL) prior with parameter vector $\vec{\Phi}$ from which the mixing weight π is generated, we compute the joint distribution of (π, z, x), the associated variational distribution, and the lower bound, respectively, as outlined below:

$$p(\pi, z, x|\vec{\Phi}, \Theta, f) = p(\pi|\vec{\Phi}) \prod_{j=1}^{m} p(z_j|\pi) p_\psi(x_j|z_j, \Theta, f_j) \tag{79}$$

$$q(\pi, z|\vec{\Phi}, \phi, f) = q(\pi|\vec{\Omega}) \prod_{j=1}^{m} q(z_j|\phi_j) \tag{80}$$

$$\mathcal{L}(\vec{\Omega}, \phi; \vec{\Phi}, \Theta) = \mathbb{E}_q[\log\ p(\pi|\vec{\Phi})] + \mathbb{E}_q[\log\ p(z|\pi)] + \mathbb{E}_q[\log\ p(x|z, \Theta)] + \mathcal{H}(q(\pi)) + \mathcal{H}(q(z)) \tag{81}$$

where, $\vec{\Omega} = (\gamma_1, ..., \gamma_K, \gamma, \lambda)$ is the Beta-Liouville parameter vector and $\phi = (\phi_1, ..., \phi_m)$ are the multinomial parameters. The BL distribution is also a member of the exponential family (Appendix A.2), thus the expected value will be obtained by computing the derivative of its cumulant function Bakhtiari and Bouguila (2016). Consequently, the corresponding variational parameters are updated as follows:

$$\phi_{(z_j, f_j)} \propto p_\psi(x_j|z_j, \Theta, f_j) \exp\left(\Psi(\gamma_{z_j}) - \Psi(\sum_{z_j=1}^{K} \gamma_{z_j}) + \Psi(\lambda) - \Psi(\gamma + \lambda)\right) \tag{82}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{83}$$

4.4 Experimental Results

To evaluate the performance of our LGD-CNB and LBL-CNB models and to compare them with LD-CNB model for each experiment, we selected different sets of data. This assessment examines how three different priors affect the Gaussian and Discrete models.

4.4.1 Gaussian Models

As mentioned earlier, Gaussian models are suitable for features with real values. Table 4.1 displays the calculated perplexities for LD-CNB-Gaussian, LGD-CNB-Gaussian, and LBL-CNB-Gaussian models across five different datasets. These datasets are chosen from the UCI benchmark repository, in which all features are available for every instance. The model was trained using 70% of the data, and the remaining 30% was utilized for testing. Perplexity values are then computed on the testing set using equation 84, with the same number of selected features for all instances in the dataset. The perplexity values after 10 iterations are presented in Table 1. According to equation 32, lower perplexity indicates a higher log-likelihood probability, suggesting a better fit for the model.

$$Perplexity(X) = \exp\Big\{-\frac{\sum_{i=1}^{N}\log p(x_i)}{\sum_{i=1}^{N} m_i}\Big\} \tag{84}$$

Table 4.1: Perplexity of LD-CNB, LGD-CNB, and LBL-CNB Gaussian models.

Dataset	LD-CNB	LGD-CNB	LBL-CNB
Wine	0.9936	0.9804	0.93262
Balance	0.9966	0.9810	0.9953
HeartFailure	0.9837	0.8792	0.7987
WDBC	0.9967	0.9943	0.9925
Yeast	0.9974	0.9963	0.9959

Results indicate that LGD-CNB-Gaussian and LBL-CNB-Gaussian models perform better than LD-CNB-Gaussian, showing that the generalized structure of GD distribution and BL distribution makes them more suitable as prior distributions. Table 4.2 displays the outcomes of assessing the models on the WDBC dataset Wolberg and Street (1995). These results represent the averages obtained from 20 runs with distinct randomly assigned initial values. According to the table, both the LGD-CNB model and LBL-CNB model outperform the LD-CNB model, showcasing higher accuracy in those instances.

Table 4.2: Accuracy, precision, recall, and f-score in percent for LD-CNB, LGD-CNB, LBL-CNB using the WDBC dataset.

Method	Accuracy	Precision	Recall	F-score
LD-CNB	0.63	0.55	0.075	0.132
LGD-CNB	0.85	0.75	0.90	0.818
LBL-CNB	0.82	0.92	0.55	0.688

4.4.2 Discrete Models

To assess Discrete models, we utilized the 100K MovieLens dataset from the Grouplens Research Project Harper and Konstan (2015). This dataset comprises $100,000$ ratings (1-5) provided by 943 users for 1682 movies. Users with fewer than 20 ratings or incomplete demographic information were excluded. Due to users not rating all movies, there is sparsity in this dataset. Similar to the prior experiment, we conducted the experiment on our three models and computed the perplexity for each using equation 86.

$$Perplexity(X) = \exp\Big\{-\frac{\sum_{i=1}^{N}\log p(x_i)}{N}\Big\} \tag{85}$$

The disparity in the number of rated movies among users serves as evidence for the dataset's sparsity, a notable distinction between the Discrete model and the Gaussian model. Moreover, there are no constraints on the covariance of data points, implying that a user rating fewer movies does not necessitate another user to rate more. As illustrated in Figure 4.1, despite the sparsity, the perplexity of the LGD-CNB-Discrete model is lower than that of the LBL-Discrete model, which, in turn, is lower than the LD-CNB-Discrete model. This finding suggests that similar to real-valued features, the more general covariance structure of the GD and BL priors allows them to better describe proportional data, leading to their superior performance over the Dirichlet prior to categorical features. Additionally, the presence of two vector parameters in the GD distribution enhances its flexibility with sparse data, enabling it to assign low values effectively Najar and Bouguila (2022c). In this experiment, we compute similarly the accuracy, precision, recall, and F-score for this model using

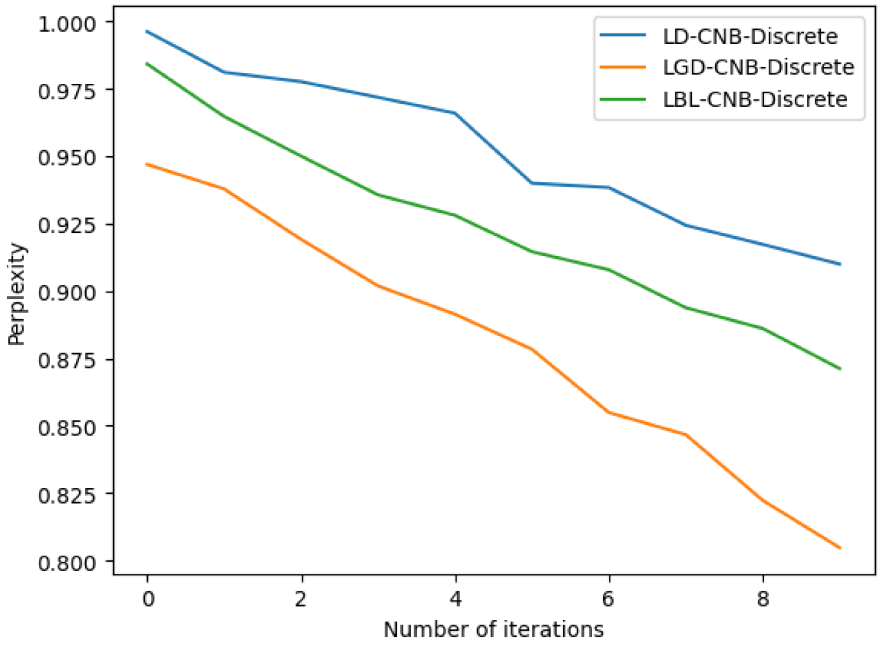


Figure 4.1: Perplexity for LD-CNB-Discrete, LGD-CNB-Discrete, and LBL-CNB-Discrete.

the testing set. Table 4.3 illustrates that the suggested approaches have led to an enhancement in the overall performance.

Table 4.3: Accuracy, precision, recall and f-score in percent for LD-CNB, LGD-CNB, LBL-CNB Discrete models.

Method	Accuracy	Precision	Recall	F-score
LD-CNB	0.83	0.62	0.052	0.095
LGD-CNB	0.87	0.75	0.51	0.607
LBL-CNB	0.86	0.74	0.39	0.51

4.5 Conclusion

In this chapter, we have presented the incorporation of the GD and BL distributions as priors in the CNB model to address sparsity in large-scale datasets. Utilizing the conditional Naive Bayes

(CNB) model, we conditioned the model on observed feature subsets, enhancing sparsity management. The traditional approach assumed a Dirichlet distribution as a prior, LD-CNB acknowledges that feature values are generated from an exponential family distribution, varying depending on the considered feature. We have outlined the advantages of employing GD and BL distributions over the Dirichlet distribution. Our investigation into the Gaussian and Discrete distributions as exponential families for LGD-CNB and LBL-CNB models revealed that the more generalized covariance structure of GD and BL distributions makes them desirable as prior distributions for uncovering latent structures in sparse data, especially when feature vectors follow a discrete distribution.

Chapter 5

Conclusion

In conclusion, this thesis has delved into the realm of Bayesian models within natural language processing, presenting a series of innovative unsupervised learning approaches designed for the identification of pertinent documents and text mining. The challenges of incorporating new queries into the topic space, addressing data sparsity, and handling positively correlated data were successfully navigated by applying various distributions within our Bayesian framework. While mixture models have been extensively explored in previous research, the selection of an appropriate prior distribution remains a critical consideration. Chapter 2 focused on developing an iterative statistical method to address the challenge of folding new queries into the topic space. This involved estimating model parameters and hidden variables, specifically topic mixtures, to establish the relevance between new queries and documents within the data collection. Moving to Chapter 3, we concentrated on the utilization of large datasets and explored the joint impact of clustering and projection in our probabilistic model. The investigation encompassed document clustering, word projection, and document modeling. In Chapter 4, our attention turned to the performance of different prior distributions in managing data sparsity. Gaussian and Discrete distributions were considered as exponential family distributions for the dataset features, forming the basis of our model. In our modeling endeavors, we opted to substitute the Dirichlet distribution with both the generalized Dirichlet distribution and the Beta-Liouville distribution within each model. This strategic choice was grounded in the favorable properties exhibited by these distributions, showcasing their proficiency in representing

proportional data. In contrast to the Dirichlet distribution, both the generalized Dirichlet distribution and the Beta-Liouville distribution have more general covariance structures. This implies the absence of negative covariance restrictions on the data, allowing for a broader range of relationships to be accommodated. Furthermore, the decision to introduce these alternative distributions comes with the added advantage of defining more independent parameters. This deliberate increase in the number of parameters injects additional degrees of freedom into our model. Consequently, this augmentation affords our model greater flexibility to adapt and offer a more precise fit to the intricacies of the data. The surplus degrees of freedom provide the model with increased adaptability, enabling it to better account for variability within the dataset. Notably, both the generalized Dirichlet distribution and the Beta-Liouville distribution possess the capacity to capture a more intricate dependence pattern among variables. This heightened flexibility allows these distributions to excel in capturing complex relationships within the data, making them valuable assets for modeling scenarios characterized by nuanced and interrelated variables. In summary, the incorporation of the generalized Dirichlet distribution and the Beta-Liouville distribution not only broadens the scope of covariance structures but also enhances the adaptability and flexibility of our models, making them more adept at handling diverse and complex datasets. Ultimately, a text mining model aims to efficiently discover and retrieve relevant information based on user queries in terms of both time and cost-effectiveness. By presenting an iterative algorithm and offering a flexible choice of prior distribution, Bayesian models emerge as valuable tools in text mining. The contributions made in this thesis pave the way for enhanced approaches in natural language processing and signify the ongoing importance of addressing delicate challenges in unsupervised learning.

Appendix A

Appendix

A.1 Exponential Form of the Generalized Dirichlet Distribution

The exponential family of distributions is a group of parametric probability distributions with specific mathematical characteristics, making them easily manageable from both statistical and mathematical perspectives. This family encompasses various distributions like normal, exponential, log-normal, gamma, chi-squared, beta, Dirichlet, Bernoulli, and more. Given a measure η, an exponential family of probability distributions is identified as distributions whose density (in relation to η) follows a general form:

$$p(x|\eta) = h(x) \exp\left(\eta^T T(x) - A(\eta)\right) \tag{86}$$

where, $h(x)$ is referred to as the base measure, $T(x)$ is the sufficient statistic. η is known as natural parameter, and $A(\eta)$ is defined as the cumulant function.

It has been shown that the generalized Dirichlet distribution is a member of the exponential family distributions Zamzami and Bouguila (2019a, 2019b, 2022), as evidenced by its representation in the aforementioned form, as illustrated below:

$$
\begin{aligned}
GD(\vec{X}|\alpha_1,...,\alpha_K,\beta_1,...,\beta_K) = \exp\Bigg[&\sum_{k=1}^{K}(\log(\Gamma(\alpha_k+\beta_k)) - \log(\Gamma(\alpha_k)) \\
&- \log(\Gamma(\beta_k))) + \alpha_1\log(X_1) + \sum_{k=2}^{K}\alpha_k\Big(\log(X_k) - \log(1-\sum_{t=1}^{k-1}X_t)\Big) \\
&+ \beta_1\log(1-X_1) + \sum_{k=2}^{K}\beta_k\Big(\log(1-\sum_{t=1}^{k}X_t) - \log(1-\sum_{t=1}^{k-1}X_t)\Big) \\
&- \sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{k=1}^{K}X_k)\Bigg]
\end{aligned}
\tag{87}
$$

Based on that we can calculate the base measure, the sufficient statistic, and the cumulant function as Bouguila (2011a); Epaillard and Bouguila (2019):

$$
h(\vec{X}) = -\sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{t=1}^{K}X_t) \tag{88}
$$

$$
\begin{aligned}
T(\vec{X}) = \Big(&\log(X_1), \log(X_2) - \log(1-X_1), \log(X_3) - \log(1-X_1-X_2), ..., \log(1-X_1), \\
&\log(1-X_1-X_2) - \log(1-X_1), ..., \log(1-\sum_{t=1}^{K}X_t) - \log(1-\sum_{t=1}^{K-1}X_t)\Big)
\end{aligned}
\tag{89}
$$

$$
A(\eta) = \Big(\sum_{k=1}^{K}\log(\Gamma(\alpha_k)) + \log(\Gamma(\beta_k)) - \log(\Gamma(\alpha_k+\beta_k))\Big) \tag{90}
$$

given $\eta = (\alpha_1, ..., \alpha_K, \beta_1, ..., \beta_K)$.

A.2 Exponential Form of the Beta-Liouville Distribution

Besides the generalized Dirichlet distribution, the Beta-Liouville distribution can also be expressed in the framework of exponential family distributions as it is shown below:

$$BL(\vec{X}|\alpha_1,...,\alpha_K,\alpha,\beta) = \exp\Bigg[\log(\Gamma(\sum_{k=1}^{K}\alpha_k)) - \log(\Gamma(\alpha+\beta)) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) \\ - \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) - \log(\Gamma(\beta)) - \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) + \sum_{k=1}^{K}\alpha_k\Big(\log(X_K) - \log(\sum_{k=1}^{K}X_k)\Big) \\ + \alpha\log(\sum_{k=1}^{K}X_k) + \beta\log(1-\sum_{k=1}^{K}X_k) - \sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{k=1}^{K}X_k)\Bigg] \tag{91}$$

In this scenario, the determination of the base measure, sufficient statistic, and cumulant function is carried out as follows:

$$h(\vec{X}) = -\sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{t=1}^{K}X_t) \tag{92}$$

$$T(\vec{X}) = \Big(\log(X_1) - \log(\sum_{k=1}^{K}X_k), \log(X_2) - \log(\sum_{k=1}^{K}X_k), ..., \log(X_K) - \log(\sum_{k=1}^{K}X_k)\Big) \tag{93}$$

$$A(\eta) = \log\Big(\Gamma(\sum_{k=1}^{K}\alpha_k)\Big) + \log(\Gamma(\alpha+\beta)) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) - \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) \tag{94}$$

given $\eta = (\alpha_1,...,\alpha_K,\alpha,\beta)$.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Bakhtiari, A. S., & Bouguila, N. (2011). An expandable hierarchical statistical framework for count data modeling and its application to object classification. In *2011 ieee 23rd international conference on tools with artificial intelligence* (pp. 817–824).

Bakhtiari, A. S., & Bouguila, N. (2012). A novel hierarchical statistical model for count data modeling and its application in image classification. In T. Huang, Z. Zeng, C. Li, & C. Leung (Eds.), *Neural information processing - 19th international conference, ICONIP 2012, doha, qatar, november 12-15, 2012, proceedings, part II* (Vol. 7664, pp. 332–340). Springer.

Bakhtiari, A. S., & Bouguila, N. (2014). A variational bayes model for count data learning and classification. *Engineering Applications of Artificial Intelligence*, *35*, 176–186.

Bakhtiari, A. S., & Bouguila, N. (2016). A latent beta-liouville allocation model. *Expert Systems with Applications*, *45*, 260–272.

Banerjee, A., & Shan, H. (2007). Latent dirichlet conditional naive-bayes models. In *Seventh ieee international conference on data mining (icdm 2007)* (pp. 421–426).

Berrar, D. (2018). Bayes' theorem and naive bayes classifier. *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, *403*, 412.

Blei, D., Ng, A., & Jordan, M. (2001). Latent dirichlet allocation. *Advances in neural information processing systems*, *14*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bouguila, N. (2007). Spatial color image databases summarization. In *2007 ieee international conference on acoustics, speech and signal processing-icassp'07* (Vol. 1, pp. I–953).

Bouguila, N. (2008). Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, *20*(4), 462–474.

Bouguila, N. (2009). A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. *IEEE Transactions on Knowledge and Data Engineering*, *21*(12), 1649–1664.

Bouguila, N. (2010). Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, *22*(2), 186–198.

Bouguila, N. (2011a). Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Transactions on Knowledge and Data Engineering*, *24*(12), 2184–2202.

Bouguila, N. (2011b). Texture discrimination using local features and count data models. In *2011 international conference on communications, computing and control applications (ccca)* (pp. 1–6).

Bouguila, N. (2012). Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognition Letters*, *33*(2), 103–110.

Bouguila, N., & ElGuebaly, W. (2008a). A generative model for spatial color image databases categorization. In *2008 ieee international conference on acoustics, speech and signal processing* (pp. 821–824).

Bouguila, N., & ElGuebaly, W. (2008b). On discrete data clustering. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 503–510).

Bouguila, N., & ElGuebaly, W. (2009). Discrete data clustering using finite mixture models. *Pattern Recognition*, *42*(1), 33–42.

Bouguila, N., & Ghimire, M. N. (2010). Discrete visual features modeling via leave-one-out likelihood estimation and applications. *Journal of Visual Communication and Image Representation*, *21*(7), 613–626.

Bouguila, N., & Ziou, D. (2004a). Improving content based image retrieval systems using finite multinomial dirichlet mixture. In *Proceedings of the 2004 14th ieee signal processing society*

workshop machine learning for signal processing, 2004. (pp. 23–32).

Bouguila, N., & Ziou, D. (2004b). A powerful finite mixture model based on the generalized dirichlet distribution: unsupervised learning and applications. In *Proceedings of the 17th international conference on pattern recognition, 2004. icpr 2004.* (Vol. 1, pp. 280–283).

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

Bouguila, N., & Ziou, D. (2009). A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, *21*(1), 107–122.

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on information and knowledge management* (pp. 211–218).

Caballero, K. L., Barajas, J., & Akella, R. (2012). The generalized dirichlet distribution in enhanced topic detection. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 773–782).

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, *163*, 1–13.

Connor, R. J., & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, *64*(325), 194–206.

Daghyani, M., Zamzami, N., & Bouguila, N. (2019). Efficient computation of log-likelihood function in clustering overdispersed count data using multinomial beta-liouville distribution. In *2019 ieee symposium series on computational intelligence (ssci)* (pp. 986–993).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, *29*, 103–130.

Dumais, S. T., et al. (2004). Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.*, *38*(1), 188–230.

Eita, A. Y., Ennajari, H., & Bouguila, N. (2023). 3d multi-views object classification based on a fully generalized dirichlet allocation model. In *2023 ieee international conference on industrial technology (icit)* (pp. 1–7).

Epaillard, E., & Bouguila, N. (2019). Data-free metrics for dirichlet and generalized dirichlet mixture-based hmms–a practical study. *Pattern Recognition*, *85*, 207–219.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).

Fan, W., & Bouguila, N. (2013). Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE transactions on neural networks and learning systems*, *24*(11), 1850–1862.

Fan, W., & Bouguila, N. (2015). Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Engineering Applications of Artificial Intelligence*, *43*, 1–14.

Fang, K.-T., Kotz, S., & Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Frank, A. (2010). Uci machine learning repository. *http://archive. ics. uci. edu/ml*.

Fu, Q., & Banerjee, A. (2008). Multiplicative mixture models for overlapping clustering. In *2008 eighth ieee international conference on data mining* (pp. 791–796).

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing*, *2*(2), 291–298.

Gopalan, R., & Berry, D. A. (1998). Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association*, *93*(443), 1130–1139.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl_1), 5228–5235.

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1–19.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hinneburg, A., Gabriel, H.-H., & Gohr, A. (2007). Bayesian folding-in with dirichlet kernels for plsi. In *Seventh ieee international conference on data mining (icdm 2007)* (pp. 499–504).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval* (pp. 50–57).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*, 177–196.

Hofmann, T., Puzicha, J., & Jordan, M. (1998). Learning from dyadic data. *Advances in neural information processing systems*, *11*.

Huang, R., Yu, G., Wang, Z., Zhang, J., & Shi, L. (2012). Dirichlet process mixture model for document clustering with feature partition. *IEEE Transactions on knowledge and data engineering*, *25*(8), 1748–1759.

Ji, X., & Xu, W. (2006). Document clustering with prior knowledge. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (pp. 405–412).

Jin, X., Zhou, Y., & Mobasher, B. (2004). Web usage mining based on probabilistic latent semantic analysis. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 197–205).

Koochemeshkian, P., Zamzami, N., & Bouguila, N. (2020). Flexible distribution-based regression models for count data: Application to medical diagnosis. *Cybern. Syst.*, *51*(4), 442–466.

Kyoya, S., & Yamanishi, K. (2021). Summarizing finite mixture model with overlapping quantification. *Entropy*, *23*(11), 1503.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Li, T., & Ma, J. (2023). Dirichlet process mixture of gaussian process functional regressions and its variational em algorithm. *Pattern Recognition*, *134*, 109129.

Li, W., & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic

correlations. In *Proceedings of the 23rd international conference on machine learning* (pp. 577–584).

Li, X., Ling, C. X., & Wang, H. (2016). The convergence behavior of naive bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(1), 1–24.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309–317.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159–165.

Luo, Z., Amayri, M., Fan, W., & Bouguila, N. (2023). Cross-collection latent Beta-Liouville allocation model training with privacy protection and applications. *Applied Intelligence*, *53*(14), 17824–17848. Retrieved from https://doi.org/10.1007/s10489-022-04378-3 doi: 10.1007/s10489-022-04378-3

Minka, T. P., & Lafferty, J. (2012). Expectation-propogation for the generative aspect model. *arXiv preprint arXiv:1301.0588*.

Najar, F., & Bouguila, N. (2021a). Smoothed generalized dirichlet: A novel count-data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, *3*(5), 685–698.

Najar, F., & Bouguila, N. (2021b). Sparse document analysis using beta-liouville naive bayes with vocabulary knowledge. In *International conference on document analysis and recognition* (pp. 351–363).

Najar, F., & Bouguila, N. (2022a). Emotion recognition: A smoothed dirichlet multinomial solution. *Engineering Applications of Artificial Intelligence*, *107*, 104542.

Najar, F., & Bouguila, N. (2022b). Exact fisher information of generalized dirichlet multinomial distribution for count data modeling. *Information Sciences*, *586*, 688–703.

Najar, F., & Bouguila, N. (2022c). Sparse generalized dirichlet prior based bayesian multinomial estimation. In *Advanced data mining and applications: 17th international conference, adma 2021, sydney, nsw, australia, february 2–4, 2022, proceedings, part ii* (pp. 177–191).

Nguyen, D. A., Nguyen, K. A., Nguyen, C. H., Than, K., et al. (2021). Boosting prior knowledge in streaming variational bayes. *Neurocomputing*, *424*, 143–159.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of*

mathematical statistics, *33*(3), 1065–1076.

Reynolds, A. P., Richards, G., & Rayward-Smith, V. J. (2004). The application of k-medoids and pam to the clustering of rules. In *International conference on intelligent data engineering and automated learning* (pp. 173–178).

Ribeiro-Neto, B., Silva, I., & Muntz, R. (2000). Bayesian network models for information retrieval. *Soft Computing in Information Retrieval: techniques and applications*, 259–291.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, 832–837.

Sahami, M., Hearst, M., & Saund, E. (1996). Applying the multiple cause mixture model to text categorization. In *Icml* (Vol. 96, pp. 435–443).

Şahin, B., Evren, A. A., Tuna, E., Şahinbaşoğlu, Z. Z., & Ustaoğlu, E. (2023). Parameter estimation of the dirichlet distribution based on entropy. *Axioms*, *12*(10), 947.

Saul, L., & Pereira, F. (1997). Aggregate and mixed-order markov models for statistical language processing. *arXiv preprint cmp-lg/9706007*.

Smucker, M. D., & Allan, J. (2005). *An investigation of dirichlet prior smoothing's performance advantage* (Tech. Rep.). Citeseer.

Taheri, S., Mammadov, M., & Bagirov, A. M. (2010). Improving naive bayes classifier using conditional probabilities.

Verbeek, J., Vlassis, N., & Nunnink, J. (2003). A variational em algorithm for large-scale mixture modeling. In *9th annual conference of the advanced school for computing and imaging (asci'03)* (pp. 136–143).

Wang, D., Xu, Y., Li, M., Duan, Z., Wang, C., Chen, B., ... others (2022). Knowledge-aware bayesian deep topic model. *Advances in Neural Information Processing Systems*, *35*, 14331–14344.

Weaver, W. (1952). Translation. In *Proceedings of the conference on mechanical translation.*

Wickramasinghe, I., & Kalutarage, H. (2021). Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293.

Wolberg, M. O. S. N., William, & Street, W. (1995). *Breast Cancer Wisconsin (Diagnostic).* UCI

Machine Learning Repository. (DOI: https://doi.org/10.24432/C5DW2B)

Wong, T.-T. (1998). Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, *97*(2-3), 165–181.

Wong, T.-T. (2009). Alternative prior assumptions for improving the performance of naïve bayesian classifiers. *Data Mining and Knowledge Discovery*, *18*, 183–213.

Wood, J., Tan, P., Wang, W., & Arnold, C. (2017). Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *2017 ieee 33rd international conference on data engineering (icde)* (pp. 411–422).

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust data mining* (pp. 27–33). Springer.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval* (pp. 267–273).

Yang, Y., Downey, D., & Boyd-Graber, J. (2015). Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 308–317).

Yazdi, S. S., Najar, F., & Bouguila, N. (n.d.). *Generalizing conditional naive bayes model.* (Submitted for publication to *2024 International Conference on Enterprise Information Systems (ICEIS)*)

Yazdi, S. S., Najar, F., & Bouguila, N. (2022). Bayesian folding-in using generalized dirichlet and beta-liouville kernels for information retrieval. In *2022 ieee symposium series on computational intelligence (ssci)* (pp. 1430–1435).

Yazdi, S. S., Najar, F., & Bouguila, N. (2023). Generalized probabilistic clustering projection models for discrete data. In *2023 international symposium on networks, computers and communications (isncc)* (pp. 1–7).

Yu, S., Yu, K., Tresp, V., & Kriegel, H.-P. (2005). A probabilistic clustering-projection model for discrete data. In *European conference on principles of data mining and knowledge discovery* (pp. 417–428).

Zamzami, N., & Bouguila, N. (2019a). Model selection and application to high-dimensional count

data clustering - via finite EDCM mixture models. *Appl. Intell.*, *49*(4), 1467–1488.

Zamzami, N., & Bouguila, N. (2019b). A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognit.*, *95*, 36–47.

Zamzami, N., & Bouguila, N. (2020a). High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution. *Information Sciences*, *524*, 116–135.

Zamzami, N., & Bouguila, N. (2020b). High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution. *Inf. Sci.*, *524*, 116–135.

Zamzami, N., & Bouguila, N. (2022). Sparse count data clustering using an exponential approximation to generalized dirichlet multinomial distributions. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(1), 89-102.