Extensions to the Latent Dirichlet Allocation Topic Model Using Flexible Priors

Koffi Eddy Ihou

A Thesis
in
Concordia Institute for Information Systems Engineering (CIISE)

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

March 2021

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Koffi Eddy Ihou**
Entitled: **Extensions to the Latent Dirichlet Allocation Topic Model Using Flexible Priors**

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

Doctor of Philosophy (Information Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.
Signed by the Final Examining Committee:

Supervisor
Dr. Nizar Bouguila

Co-supervisor
Dr. Wassim Bouachir

Examiner
Dr. Abdessamad Ben Hamza

Examiner
Dr. Arash Mohammadi

Examiner
Dr. Zachary Patterson

External Examiner
Dr. Benjamin Fung

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

2021

Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Extensions to the Latent Dirichlet Allocation Topic Model Using Flexible Priors

Koffi Eddy Ihou, Ph.D.
Concordia University, 2021

Intrinsically, topic models have always their likelihood functions fixed to multinomial distributions as they operate on count data instead of Gaussian data. As a result, their performances ultimately depend on the flexibility of the chosen prior distributions when following the Bayesian paradigm compared to classical approaches such as PLSA (probabilistic latent semantic analysis), unigrams and mixture of unigrams that do not use prior information. The standard LDA (latent Dirichlet allocation) topic model operates with symmetric Dirichlet distribution (as a conjugate prior) which has been found to carry some limitations due to its independent structure that tends to hinder performance for instance in topic correlation including positively correlated data processing. Compared to classical ML estimators, the use of priors ultimately presents another unique advantage of smoothing out the multinomials while enhancing predictive topic models.

In this thesis, we propose a series of flexible priors such as generalized Dirichlet (GD) and Beta-Liouville (BL) for our topic models within the collapsed representation, leading to much improved CVB (collapsed variational Bayes) update equations compared to ones from the standard LDA. This is because the flexibility of these priors improves significantly the lower bounds in the corresponding CVB algorithms. We also show the robustness of our proposed CVB inferences when using simultaneously the BL and GD in hybrid generative-discriminative models where the generative stage produces good and heterogeneous topic features that are used in the discriminative stage by powerful classifiers such as SVMs (support vector machines) as we propose efficient probabilistic kernels to facilitate processing (classification) of documents based on topic signatures. Doing so, we implicitly cast topic modeling which is an unsupervised learning method into a supervised learning technique.

Furthermore, due to the complexity of the CVB algorithm (as it requires second order Taylor expansions) in general, despite its flexibility, we propose a much simpler and tractable update equation using a MAP (maximum a posteriori) framework with the standard EM (expectation-maximization) algorithm. As most Bayesian posteriors are not tractable for complex models, we ultimately propose the MAP-LBLA (latent BL allocation) where we characterize the contributions of asymmetric BL priors over the symmetric Dirichlet (Dir). The proposed MAP technique importantly offers a point estimate (mode) with a much tractable solution. In the MAP, we show that point estimate could be easy to implement than full Bayesian analysis that integrates over the entire parameter space. The MAP implicitly exhibits some equivalent relationship with the CVB especially the zero order approximations CVB0 and its stochastic version SCVB0. The proposed method enhances performances in information retrieval in text document analysis.

We show that parametric topic models (as they are finite dimensional methods) have a much smaller hypothesis space and they generally suffer from model selection. We therefore propose a Bayesian nonparametric (BNP) technique that uses the Hierarchical Dirichlet process (HDP) as conjugate prior to the document multinomial distributions where the asymmetric BL serves as a diffuse (probability) base measure that provides the global atoms (topics) that are shared among documents. The heterogeneity in the topic structure helps in providing an alternative to model selection because the nonparametric topic model (which is infinite dimensional with a much bigger hypothesis space) could now prune out irrelevant topics based on the associated probability masses to only retain the most relevant ones.

We also show that for large scale applications, stochastic optimizations using natural gradients of the objective functions have demonstrated significant performances when we learn rapidly both data and parameters in online fashion (streaming). We use both predictive likelihood and perplexity as evaluation methods to assess the robustness of our proposed topic models as we ultimately refer to probability as a way to quantify uncertainty in our Bayesian framework. We improve object categorization in terms of inferences through the flexibility of our prior distributions in the collapsed space. We also improve information retrieval technique with the MAP and the HDP-LBLA topic models while extending the standard LDA. These two applications present the ultimate capability of enhancing a search engine based on topic models.

Acknowledgments

I would like to sincerely thank Dr. Nizar Bouguila, my academic advisor for my PhD program at Concordia University. Throughout these years, his support, encouragement, advice, and love for research have all been instrumental and valuable in the development of this work (thesis). I would like to thank him again and again for being such an exceptional advisor to me in so many aspects.

I would also like to thank Dr. Wassim Bouachir, my co-advisor, for his undeniable support, contribution, and advice in my academic life as a PhD student. He has always shown his availability to help.

As a result, I do owe my deepest gratitude to Dr. Nizar Bouguila and Dr. Wassim Bouachir, not only for accepting me, but also for being part of this journey of mine as a PhD student and a researcher.

I would like to take this opportunity to thank the members of my Defence Committee, for their presence and contributions during my Defence session. The comments, suggestions, advice, and encouragements I received have all been valuable to me.

Ultimately, I would like to thank Concordia University and NSERC (Natural Sciences and Engineering Research Council of Canada) for supporting my research.

Finally, I would like to thank my dear parents and sisters for their unconditional love and support. Their encouragement has always been meaningful to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With advancement in internet technology, proliferation of hardware (printers and scanners, mobile phones and cameras), and the development of social media platforms, our 21st century society continues to collect unprecended amount of information for large scale applications. Processing such unstructured records requires efficient machine learning techniques due to the complexity and variability in these massive collections (images, text documents, 3D objects, videos, and their combinations). In topic modeling, such collections are summarized as documents that operate with count data following the BoW (bag of words) method. The goal is to build good topics in order to make efficient prediction on unseen documents in tasks such as retrieval and classification. The topics represent the intermediate low dimensional (subspace) representation of documents [2], [3]. The widely known topic model is the standard LDA with its Dirichlet distribution as conjugate prior to the multinomial. In LDA, documents arise as mixture over topics while the topics are distributions over the vocabulary words. The LDA has been implemented as a direct alternative to the frequentist method (classical maximum likelihood estimate approaches) because of its ability to smooth out multinomials using Dirichlet. Classical frequentist methods do not use prior informations, and this complicates their performances when it comes to predicting previously unseen documents or events. Compared to LDA, the unigram model draws words in a document from a single multinomial distribution called word simplex. The mixture of unigrams is an augmented version of the unigram model with a discrete topic (latent) variable. The mixture of unigrams, generates a document from only a single topic [3]. The PLSA is almost identical to LDA topic model, however, it has no prior information [4]. It relaxes the mixture of unigrams assumption as it allows a document to exhibit multiple topics. As presented earlier, the lack of priors in PLSA makes the model unfit for prediction and often suffers from overfitting problems. The LDA topic model provides a solution to the PLSA, unigram, and mixture of unigrams by including prior information as it treats topic proportions as random variables [5].

Topic models therefore depend extensively on prior information because their likelihood functions are fixed to multinomial distributions, so their robustness ultimately depends on the use of flexible priors. In fact multinomial has some limitations in topic modeling: for instance, using only frequencies as ways to represent probabilities often leads to poor estimates. In a highly sparse collection, without any smoothing method, frequencies are more likely to assign zero probabilities for unseen or rare events. Moreover, and very often, multinomials do not capture very well the words burstiness because of the lack of priors [6, 7]. The integration of prior information has become fundamental for the flexibility of topic models such as LDA over classical frequentist approaches. In other words, the limitations of classic frequentist models led to the emergence of LDA and its variants.

Due to the limitations of the Dirichlet prior, we are able to reformulate the generative process with flexible priors such as GD and BL as alternatives. It is noteworthy that under the Dirichlet, topic components are independent which prohibit topic correlation framework. Because of this handicap, the LDA could not provide a natural way of organizing documents as it does not allow any dependency between topics [8]: in a collection, in a real life scenario, it is natural to observe that the existence of one topic is correlated to another one within a document or another document. This structure facilitates grouping and compression methods. While conjugate priors have been used for their simplicity in providing closed form posteriors (for exponential family distributions), some topic modeling techniques have encouraged the use of non conjugate priors as alternatives when dealing with topic correlation for instance. Asymmetric and symmetric properties have been also added to prior information for enhancing estimation of the parameters [9, 10]. As a ultimate goal, a robust and efficient topic model could be embedded into a search engine for object categorization and information retrieval.

For classification, for instance, topics learnt in the generative stage could be used in the discriminative stage with powerful classifiers such as SVM (support vector machines)[11]. This setting requires the use of machine learning techniques or inferences such as VB (variational Bayes), CVB (collapsed variational Bayes), and CGS (collapsed Gibss Sampling), and EP (expectation propagation). The CVB is a hybrid inference between the VB (deterministic) and CGS (stochastic approach) using MCMC (Markov Chain Monte Carlo)[12]. The CVB has been considered one of the state-of-the-art methods in topic modeling; nevertheless, the CVB is very complex, and computational expensive. It led to its zero order approximations CVB0 instead which is a much simpler model, but it is not efficient in large scale applications within parametric topic modeling because of the reduced hypthesis space. The stochastic CVB0 (SCVB0) is the online version of the CVB0 [13]. Still in parametric model, since most Bayesian posteriors, for complex models, are intractable in general, a point estimate (the mode) offers a much tractable solution. The MAP hypotheses using point estimates are much easier than full Bayesian analysis that integrates over the entire parameter space. The MAP could reduce the three-level hierarchical LDA to two level topic mixture as it marginalizes out the latent variables leaving the parameters. The MAP in contrast to CVB integrates out the parameters leaving the latent variables. In addition to the limitation of parametric models in model selection due to their reduced hypothesis space, the Bayesian nonparametric framework has been implemented in LDA (HDP-LDA) [14, 15, 16]. Using the BL and GD priors we are able to derive a variety of parametric topic models before implementing a nonparametric method which solves three main problems in topic modeling: extending the LDA capabilities, re-assuring the sharing ability of clusters within or across groups (documents), and the model selection ability. Ultimately, under the nonparametric setting, the data choose the number of topics by themeselves.

We can summarize our intentions by confirming that we formulate this thesis based on the observation that most of the traditional inferences in topic models, parametric and nonparametric along with their stochastic approximations only deal with the LDA which extensively uses its Dirichlet prior. Now with the limitation of Dirichlet and the emergence of flexible (conjugate) priors that generalize the LDA, it became natural and straightforward for us to extend the capabilities of the LDA. Furthermore, most of these LDA-based models became very restrictive in performance in large scale applications [17]. In this thesis, we reformulate the collapsed representation (in chapter 2) using the generalized Dirichlet (GD) priors as alternatives to the Dirichlet distributions followed by another inference in the collapsed space (in chapter 3) with the Beta-Liouville (BL) priors. We also presented a hybrid generative-discriminative model (in chapter 4) that uses topic features in the generative stage for a classification framework in the discriminative stage with SVMs. The proposed generative stage generates topics by utilizing the GD and BL simultaneously. We characterize efficient probabilistic kernels to accommodate the classification process. Though, to simplify the complexity of the CVB algorithms, we therefore propose (in chapter 5) the MAP method where we implement the BL prior leading to an EM lower bound (very simple EM lower bound) with the the MAP-LBLA as alternative to the MAP-LDA. Finally, to improve performances in standard nonparametric topic models with LDA that widely use symmetric prior (Dirichlet), we propose (in chapter 6) an efficient Bayesian nonparametric technique that enhances the HDP (hierarchical Dirichlet process)'s ability to model selection framework with the HDP-LBLA topic model as we utilize the BL as a diffuse base measure. The proposed method highly increases even further the possibility of sharing more topics (clusters) between documents. These contributions ultimately summarize this thesis work which is going to be more elaborated in the following section 1.1

1.1 Thesis Overview

This thesis is structured as follows:

□ Chapter 1 presents the intrinsic properties of topic models including the standard LDA and the classical approaches such as PLSI also called PLSA, unigrams and mixture of unigrams. Topic models have their likelihood set to multinomial distribution so their robustness depends on their ability to carry conjugate flexible priors following the Bayesian paradigm for accuracy in the estimates. In the next we have carried out new inferences using flexible priors such as GD abd BL.

□ Chapter 2 focuses on the collapsed representation. It is an important method in topic modeling as it provides a much robust lower bound for the variational framework. We extend the collapsed variational update equation using asymmetric GD prior as an alternative to the standard symmetric Dirichlet prior widely used in LDA. We therefore developed a collapsed variational inference for the latent generalized allocation topic model which extends the LDA architecture in a collapsed variational Bayes setting. The predictive models were recorded to be more accurate. Due to the ability to characterize dependency between latent variables and model paramters (hidden variables), the predictive models are much accurate with an easy access to model selection: within parametric setting, we choose the number of clusters (topics) and vocabulary size based on the probability mass functions.

□ Chapter 3 also presents another extension to the collapsed variational update equation using asymmetric BL prior where we characterize dependency in the hidden variables while also discussing about model selection similar to chapter 2 (we realized that the BL and GD gerenalize the Dir prior. However, the GD has twice the number of parameters of Dirichlet while the BL has just two more parameters than the Dir). This makes the BL the more versatile prior with less number of parameters.

□ Chapter 4 introduces a hybrid model (still in the collapsed space) where the generative stage produces good topic features that are fed to discriminative classifiers (SVM). We presented very specialized probability kernels that accommodate classification framework using our generative discriminative model. To enhance the heterogeneity in the topic features, we combine the flexibility of both the GD abd BL priors as they are simultaneously used during the generative stage which implements a CVB algorithm (that also uses both distinct priors). With these two priors we obtain the GD-BL and BL-GD-based topic models whose topics are used during the discriminative stage.

□ Chapter 5 shows that the CVB inference is a very complex approach that requires second order Taylor expansion that includes mean and variance correction factor. Due to its complexity it could be intractable. The zero order approximation CVB0 has been proposed for fast batch processing along with its stochastic version (SCVB0). Due to this complexity of the CVB update equation we propose an alternative within the MAP framework for LBLA topic model using standard EM algorithm. We ultimately show that MAP-LBLA has some equivalence relationship with the CVB with LDA. The work ultimately shows that the MAP-LBLA simplifies the original CVB-LBLA update equation which facilitates the proposed stochastic optimizations using minibatches in large scale applications that require data and parameter streaming. Using the stochastic method, the time and memory complexities are much improved compared to the standard batch methods.

□ Chapter 6 covers the efficiency of variational inference using HDP prior for the document parameter in topic modeling. We propose the BL prior as the diffuse base measure which provides the global topics that are shared among documents. The BL also as a conjugate prior to the document multinomial distribution facilitates inference. We implement the HDP-LBLA in a stochastic variational inference as a direct alternative to the HDP-LDA topic model (based on symmetric Dirichlet prior). In this framework our datasets efficiently select their underlined number of topics (as alternative to models selection). Due to reduced number of topics and vocabulary size the online methods with HDP-LBLA have a much refined time and memory complexities which allows them to operate efficiently in large scale applications.

□ Chapter 7 provides a conclusion of the thesis by rearticulating the main contributions while presenting exciting research opportunities as future work.

1.2 Contributions

The main contributions of this thesis could also be summarized as follows:

□ We ultimately improved the collapsed variational Bayesian inference for LDA. This includes the batch-based CVB and the stochastic SCVB along with their zero order approximations CVB0 and SCVB0. These update equations extend the LDA topic model in the collpased space.

□ We improve the parameter estimation in exact fashion with the BL and GD priors within the collapsed representation as we relax the independent assumption in the mean-field variational inference.

□ We show that hybrid generative discriminative models could enhance performance in classification. They also cast the topic modeling method into a supervised setting when coupled with SVM. This was the case with our hybrid models using both GD and BL priors at the generative stage and the SVM at the discriminative stage with powerful and carefully selected probabilistic kernels.

□ We simplify the CVB update equation using the MAP estimation as alternative to the very complex CVB algorithm.

□ Within the Bayesian paradigm, we also showed that because the BL has few parameters compared to GD, empirical Bayes framework (hyperparameter estimation) could be faster with BL-based topic models than GD-based topic models.

□ Due to the very reduced hypothesis space of parametric topic models such as LDA, LBLA and LGDA, we propose the HDP prior for Bayesian nonparameteric topic model using asymmetric BL prior as a diffuse base measure. Using asymmetric BL, the heterogeneity in the topic stucture provides alternative to model selection as we associate each topic to its corresponding probability mass. We can therefore prune out irrelevant topics based on their weigths.

□ We show the effectiveness of the proposed stochastic optimizations as we improve time and memory complexities by reducing size in the vocabulary and number of topics in a minibatch framework.

1.3 Contributions from Authors

This PhD thesis ultimately consists of five manuscripts that represent five journal papers. Each journal (manuscript) summarizes a thesis chapter. Three journal papers have been accepted and published while the two remaining are recently submitted for publications as seen below:
First Manuscript (Chapter 2) K. E. Ihou and N. Bouguila, *Variational-based latent generalized Dirichlet allocation model in the collapsed space and applications*, Neurocomputing 332 (2019) 372-395.

Second Manuscript (Chapter 3) K. E. Ihou and N. Bouguila, *Stochastic topic models for large scale and nonstationary data*, Engineering Applications of Artificial Intelligence, volume 88, Elsevier, 2020

Third Manuscript (Chapter 4) K. E. Ihou, N. Bouguila, and W. Bouachir, *Efficient integration of generative topic models into discriminative classifiers using robust probabilistic kernels*, Pattern Analysis and Applications (2020) 1-25

Fourth Manuscript (Chapter 5) K. E. Ihou, N. Bouguila, and M. Amayri. *A Two-Level Hierarchical Latent Beta-Liouville Allocation for Large Scale Data and Parameters Streaming.* IEEE Transactions on Neural Networks and Learning Systems, (submitted for publication in 2020)

Fifth Manuscript (Chapter 6) K. E. Ihou, N. Bouguila, and M. Amayri *Stochastic Variational Optimization of A Hierarchical Dirichlet Process Latent Beta-Liouville Topic Model*. ACM Transactions on Knowledge Discovery from Data, (submitted for publication in 2020)

Chapter 2

Variational-based Latent Generalized Dirichlet Allocation Model in the Collapsed Space and Applications

In topic modeling framework, many Dirichlet-based models performances have been hindered by the limitations of the conjugate prior. It led to models with more flexible priors, such as the generalized Dirichlet distribution, that tend to capture semantic relationships between topics (topic correlation). Now these extensions also suffer from incomplete generative processes that complicate performances in traditional inferences such as VB (Variational Bayes) and CGS (Collaspsed Gibbs Sampling). As a result, the new approach, the CVB-LGDA (Collapsed Variational Bayesian inference for the Latent Generalized Dirichlet Allocation) presents a scheme that integrates a complete generative process to a robust inference technique for topic correlation and codebook analysis. Its performance in image classification, facial expression recognition, 3D objects categorization, and action recognition in videos shows its merits.

2.1 Introduction

The importance of topic modeling has drawn the attention of many researchers with exponential emergence of data from different sources. In the past, many applications have seen an extensive use of Gaussian distributions within a variety of statistical and learning frameworks. However, the inability of the Gaussian to perform effectively with count data led to the consideration of topic models such as LDA. The introduction of the LDA [3] and especially its major success in the field of topic modeling have demonstrated the early capabilities of the model. Its traditional inference schemes ranged from variational Bayes (VB) to MCMC (Markov chain Monte Carlo) approaches such as the Gibbs sampler (GS) and the collapsed Gibbs sampler (CGS) [3, 2, 18]. Topic modeling techniques have been used in a variety of applications, and ultimately led to several extensions of the LDA model. Facing storage issues and computational speed, LDA has quickly shown its ability to summarize database contents into their most relevant topics while still maintaining the intrinsic statistical structure in the database [5]. The scheme helped uncovering and maximizing the amount of information hidden behind these large collections of data.

Though, rapidly, the inability of the Dirichlet distribution to capture correlation between topics has hindered the performance of the model in several applications related to intra-class variation problems. This situation automatically forced the introduction of better, more flexible priors and models that can also guaranty the conjugacy assumption for easy Bayesian inferences. That was the case of models such as CTM (Correlated Topic Models), PAM (Pachinko Allocation Model) [8, 19, 20], IFTM (Independant Factor Topic Models) [21, 22], GD-LDA (Generalized Dirichlet-based LDA)[18], and LGDA (Latent Generalized Dirichlet Allocation) [23]. The GD-LDA for instance is an extension of the original LDA [3] that implements a generalized Dirichlet (GD) as a prior conjugate to the document multinomial distribution. It therefore replaces the Dirichlet prior in the LDA's documents modeling. Similarly, the LGDA samples the documents parameters from GD distributions. Different from the other models, the CTM utilizes the logistic normal distribution which in fact is not a conjugate prior to the multinomial distribution [8, 21]. Despite its success in topic correlation analysis, it leads to a model that is very complex and difficult to implement [8]. Consequently, in the other schemes, the introduction of the GD [24] has not only provided a very useful tool to capture correlation between topics, but also emphasized on the possibility of an easy access of the optimal number of topics (model selection). The GD mainly came as a result of the limitations of the Dirichlet distribution. Prior to the emergence of the GD, many topic modeling approaches have often used a predefined number of topics.

The ultimate goal is to prevent the model from overfitting as the database grows in size. However, with their ability to capture topic correlation, PAM and CTM are still prone to overfitting, therefore crippling these models from performing efficiently in a case where both the topic and codebook (dictionary or vocabulary) grow in size simultaneously. In addition, these two models are computationally expensive compared to the GD-LDA, CVB-LDA, LGDA, and LDA models.

Dealing with large collections of data of different types requires robust machine learning techniques that could take advantage of efficient computational methods to increase processing speed and manage data storage. One way is to construct models using efficient inference techniques as the traditional schemes are being obsolete facing the tremendous challenges and complexities of large scale datasets processing. As a result, for inferences, variational Bayes (VB) and MCMC methods, individually, are no longer the state-of-the-art inference techniques as the collapsed Gibbs sampler (CGS) is not efficient (convergence problem), and VB alone is inaccurate since it suffers from a large bias due to the strong independency assumption between latent variables and the parameters. Moreover, the relevance feedback mechanism [25, 26, 27, 28, 29] (introduced to provide an answer to the problem of optimal number of topics) using MCMC methods in IR (Information Retrieval) is computationally expensive for extremely large datasets.

The GD-LDA is designed to improve the generative process in the original (smoothed) LDA model; however, it still uses a Dirichlet prior for the vocabulary (corpus) parameter. Then, the LGDA implements the GD on document parameters while its corpus parameter was not generated to reduce computational complexities in the parameters estimation. Managing the vocabulary size is extremely important in topic modeling to avoid serious sparsity problems [3, 5]. As the vocabulary codewords influence topics estimation, a more flexible prior such as the GD for the corpus parameter could improve and effectively capture the vocabulary codewords structure (after the clustering algorithm) as it could help reducing the dictionary contents into its most relevant codewords. Due to these limitations observed in the previous models, our new approach, the CVB-LGDA improves the state-of-the-art in topic correlation framework. The CVB-LGDA model is a direct extension to the CVB-LDA. In our proppposed approach, the GD not only replaces the Dirichlet prior for the document parameter similar to the GD-LDA, but also does it for the corpus parameter. The new model in this chapter is a pure GD-based CVB model. With the shortcomings linked to the Dirichlet prior in topic correlation, the new scheme is more robust to large scale applications than the other extensions presented in this section. Its GD-based CVB algorithm also combines the advantages of its VB and CGS inferences methods for an efficient topic modeling in a scheme that favors mean field approximations, topics and vocabulary codewords analysis. Experimental results in image, 3D object categorization, and video action recognition show the generalization capabilities of the model and the LDA hierarchical architecture. One main objective of this chapter is to compare our proposed approach to the LDA and its previous extensions (variants) such as GD-LDA, LGDA, and the CVB-LDA. This, because their priors are also conjugate to the multinomials as we are maintaining this concept in our new topic model as well for easy Bayesian inference purposes. In addition, being a classification model, we are evaluating our proposed scheme and its inference technique through a comparison of its performance to other classification approaches such as BPNN (Backpropagation Neural network), SVM (Support Vector Machine), and KNN (K-Nearest Neighbor). In overall, the contribution in this new generative probabilistic model can be summarized as follows:

- The new approach provides an improvement to the generative process of the LDA [3], CVB-LDA [30], GD-LDA [18], and LGDA [23]: as large collection of data creates a large vocabulary size which often leads to a serious sparsity problem, this chapter proposes a better prior (GD) that ultimately replaces the traditional Dirichlet ditribution. It then emphasizes on smoothing the GD on the multinomial parameters (both the documents and corpus parameters). Previous models such as GD-LDA, LGDA only drew the document multinomial parameters from a GD distribution while the corpus parameters are either from Dirichlet or are not generated at all [23]. This is not efficient when dealing with datasets with a large vocabulary size.

- It directly improves the CVB-LDA. In our model, the inference is now reformulated with the GD prior, and it implements a new, robust, and complete generative process in contrast to the Dirichlet-based CVB model and other extensions using the Dirichlet prior.

- Our proposed model includes a class label to the CVB algorithm to extend the capabilities of the inference in categorization framework. It therefore represents an improvement of the CVB-LDA, LDA, LGDA, and the GD-LDA for its ability to learn its topics automatically (without human intervention) while still assigning a class label to unseen documents based on topic distribution in each class.

- The new scheme reconciles an unsupervised learning (topic modeling) to a supervised learning (classification).

The proposed approach in this chapter is structured as follows: section 4.2 illustrates the background and relative work. Section 3.3 presents the new approach while section 4.4 covers the experiments and results in several applications. Finally, section 4.5 explores some future work and provides a conclusion.

2.2 Related Work And Background

LDA [3] is a generative probabilistic model that has been introduced to solve problems in the original pLSI (probablistic Latent Semantic Indexing) [31, 32, 33]: overfitting and the difficulty in predicting documents probability outside the training set [5]. Known as a multinomial PCA (Principal Component Analysis), the LDA has especially found today its applications in text modeling and computer vision [30]. As a result, understanding all the different extensions of the LDA first necessitates a summary of the generative process in the original LDA graphical model. In this generative process of the (smoothed) LDA, documents are represented as random mixtures over the latent variables where each topic is a distribution over the vocabulary words or visual words (codewords). In this scheme, for instance, for a corpus consisting of D documents of length N_i, we usually follow these three main generative steps in the original LDA as illustrated below :
1-Choose the document parameter $\theta_i \sim Dir(\varepsilon)$ where $i \in \{1, ..., D\}$
2-Choose the corpus parameter $\varphi_k \sim Dir(\beta)$ where $k \in \{1, ..., K\}$.
3-For each word position i, j with $j \in \{1, ..., N_i\}$ and $i \in \{1, ..., D\}$
a-choose a topic $z_{ij} \sim Mult(\theta_i)$
b-choose a word $w \sim Mult(\varphi_{z_{ij}})$
such that $Mult(\theta_i)$ and $Mult(\varphi_{z_{ij}})$ are multinomial distributions with parameters θ_i and $\varphi_{z_{ij}}$, respectively, while $Dir(\varepsilon)$ and $Dir(\beta)$ are Dirichlet distributions with hyperparameters ε and β, respectively.
As observed in the LDA architecture, documents multinomial parameters θ are drawn from a Dirichlet prior with hyperparameters ε; consequently, the K-dimensional random variable θ following a Dirichlet distribution could be expressed as:

$$p(\theta|\varepsilon) = \frac{\Gamma(\sum_{k=1}^{K} \varepsilon_k)}{\prod_{k=1}^{K} \Gamma(\varepsilon_k)} \prod_{k=1}^{K} \theta_k^{\varepsilon_k - 1} \tag{1}$$

such that $\sum_{k=1}^{K} \theta_k = 1$

In the following subsections, we will discuss the major differences in the previous extensions which aim to implicitly exhibit the main contributions in our new model. Meanwhile, for the remaining of this chapter and for modeling purpose, the variables w and x could be used interchangeably to denote a codeword in an image, 3D object, and videos while the variable $\mathcal{X}$ defines a collection of x codewords within the BoW framework.

2.2.1 Differences in the generative process

Despite our approach being compared to the CVB-LDA [30], GD-LDA [18], and the LGDA [23], all these topic models follow the same generative and Bayesian hierarchical architecture of the original LDA [3, 30]. Nevertheless, each has a different generative process. Following the generative step defined above for the LDA, we can observe that in GD-LDA model [18], the document parameter is drawn from a GD distribution while the corpus parameter is still sampled from an asymmetric Dirichlet distribution. Such approach is only suitable for text modeling where the dictionary is easy to implement with the Dirichlet. Though, the performance of the model is limited when using datasets such as images and videos that require extensive topic correlation and codewords analysis. In LGDA [23], the documents parameters were also drawn from a GD distribution. However, the corpus parameter was not generated; in other words, the second step (choosing the corpus parameter) in the generative process has been avoided or neglected in the LGDA. This technique, computationally, reduces the model in the parameters estimation especially with EM (expectation-maximization) within the VB framework. However, it makes the generative process incomplete or inefficient (with the Gibbs sampler which often requires both the corpus and the document parameters to be generated) when dealing with a large vocabulary size. We might for instance want to reduce the codewords size into most relevant features or generating relevant codewords that define the documents. The CVB-LDA has the same generative model of the original and smoothed LDA with the use of the Dirichlet prior on both the document and corpus parameters. Unfortunately, this generative process is not efficient due to the limitation of the Dirichlet prior in topic correlation, and other large scale applications. In other words, the critics to the Dirichlet distribution revolve around its very restricted covariance structure that ultimately hinders its performance in topic correlation analysis since it could not be used for positively correlated data. The situation forced many of these models to operate with text datasets only as shown in [3, 18]. Moreover, all these difficulties and challenges have promoted the introduction of our new technique, the CVB-LGDA as it reformulates the generative process of the LDA where now both the corpus and documents parameters are sampled from the GD priors in the collapsed space of latent variables. The goal is to allow an effective topic and codebook analysis, and doing so makes the generative process complete, robust, efficient, and flexible for correlated topic modeling framework where both the topic and the vocabulary size could be reduced through pruning methods. This automatically improves processing (computational speed and storage) in a case of large data collections. The new extension in this proposed approach and its generative model are described in Algorithm 1 while the full comparison between the previous techniques and our model is provided by Table.2.1. Finally, the difference between these extensions can also be explained through their inference methods as shown in the next subsection.

Concerning the GD distribution, in a $(K+1)$-dimensional space, this prior with K dimensional hyperparameters $\varepsilon = (\alpha_1, \beta_1, ..., \alpha_K, \beta_K)$ is defined as:

$$p(\theta/\varepsilon) = \prod_{d=1}^{K} \frac{\Gamma\left(\alpha_d + \beta_d\right) \theta_d^{\alpha_d - 1}}{\Gamma\left(\alpha_d\right) \Gamma\left(\beta_d\right)} \left(1 - \sum_{l=1}^{d} \theta_l\right)^{\gamma_d} \qquad (2)$$

where the vector $\theta = (\theta_1, ..., \theta_K)$ is the K-dimensional multinomial parameter drawn from the GD distribution.

2.2.2 Differences in inference techniques

Before going into details in section 3.3 that is mainly dedicated to models inferences, we can briefly mention here another aspect that makes each extension different: the inferences. The lack of efficiency coupled with some other major limitations in these methods ultimately led to the implementation of our new approach. For inferences, the original LDA often uses the VB or the Gibbs sampling (MCMC) methods for the latent and parameters estimation. The Dirichlet-based CVB-LDA combines both VB and the collapsed Gibbs sampler in the collapsed space [30]. The LGDA is based on variational Bayes inference. Though, the GD-LDA favors the collapsed Gibbs sampler. The problem with the VB is that the technique suffers from a large bias as it always assumes that parameters and latent variables are independent leading to the factorization of the joint posterior distribution. This strong

	Description	Topic correlation capability
LGDA	It uses a GD prior in a VB inference, but VB alone is not always accurate. In addition, the corpus parameter is still not generated in order to simplify computations in MLE (maximum likelihood estimation).	Possible topic correlation analysis (reducing number of topics), but cannot manage the vocabulary size as the corpus (vocabulary parameter) is not generated.
GD-LDA	It implements a GD-based CGS. Though, CGS alone is also not efficient (slow and no easy access to convergence).	Possible topic correlation analysis as the documents parameters are drawn from the GD while the corpus parameter is still from a Dirichlet distribution. The model is very limited to text modeling only
LDA	It utilizes a VB or a CGS inferences. Nevertheless, it is based on the Dirichlet prior which is found to be very limited.	No topic correlation capability for positively correlated datasets due to the limitations of the Dirichlet prior.
CVB-LDA	Its CVB scheme is the current state-of-the-art, and a robust inference that combines the advantages of VB and CGS. However, it is a Dirichlet-based model (as a result, it is very limited).	Good inference technique, but no topic correlation ability for positively correlated datasets because of the Dirichlet prior.
CVB-LGDA	It is our proposed model to fix the CVB-LDA. It automatically combines the advantages of both GD based-CGS and GD based-VB inferences.	Very flexible model for correlation between topics with the GD prior. Both the topics and vocabulary codewords could be analyzed. The model is also flexible to data of different types.

Table 2.1: Comparison between the new CVB-LGDA model and the other schemes within the BoW framework

Algorithm 1 GD-based Generative Model

```
procedure
    for topic $k \leftarrow 1$ to $K$ do
        draw $\varphi_k \sim GD(\zeta)$
    end for
    for document $j \leftarrow 1$ to $D$ do
        draw $\theta_j \sim GD(\varepsilon)$
        for word $w \leftarrow 1$ to $N_j$ do
            draw $z_{wj} \sim Mult(\theta_j)$
            draw $w|z_{wj} \sim Mult(\varphi_k)$
        end for
    end for
end procedure
```

assumption could have a negative effect on the lower bound, the likelihood distribution, and the overall performance of the model when there is for instance any dependence between the parameters and latent variables. As the VB alone could be inaccurate, the Gibbs sampler (MCMC) often suffers from convergence problems [30]. Finally, the use of the Dirichlet prior in CVB-LDA approach limited its performance and hindered its ability to capture correlation between topics. First presented as a solution to VB and CGS individual drawbacks, the CVB-LDA is now inefficient and also needs a replacement due to the Dirichlet. We could observe from these inference approaches that each of the previous extensions has some limitations; therefore, there is a need for an improvement in these models. Our new method, combining both the advantages of VB and CGS with the GD as a prior solves the problem related to the Dirichlet distribution in the CVB-LDA, the original LDA, and other extensions. Furthermore, as the new approach is used in a classification problem, a category level (label) is automatically added to the hierarchical structure, as illustrated in Fig. 2.1 similar to [2]. Consequently, it improves the current CVB algorithm for classification. Overall, the new technique with its flexible priors and a robust inference technique is an extension to the LDA [30].

2.2.3 Previous Fei-Fei Li et al. 's work in image classification using topic model

The LDA model has witnessed so many extensions ultimately due to some major limitations in the model's prior (Dirichlet distribution). One of these weaknesses is the inability of the Dirichlet prior to perform in a topic correlation analysis. This, because the Dirichlet has a very limited covariance structure. In [2], despite their model providing a better way to label topics (intermediate representations) when using unsupervised learning with the LDA, the authors quickly suggested that the topic-based classification model they implemented was far from complete. In other words, the model, even though suitable for classification, was very limited: it was only successful for inter class variation problem. The scheme was not able to perform well in intra class variation problem as it could not make any difference between classes that carry almost similar features (topics) while for categories that have very distinct features there was no problem. Consequently, facing this handicap, as future work, they suggested to focus on ways that could generate richer features in order to be successful in their proposed image categorization scheme using topics. Authors Fei-Fei Li et al. in [2] did not doubt about the stability of the Dirichlet prior as they thought they could solve the classifiction problem (using images) through a preprocessing technique that could generate robust features.

Once the limitations of the Dirichlet prior have been discovered later on, scientists then decided to find a new prior (or alternative) for the LDA topic model [34, 35]. This has led to so many extensions in the quest of providing the model with the best prior. Another aspect to consider in the LDA model for classification proposed by Fei-Fei Li et al. in [2] was the inference as the variational Bayes EM (Expectation Maximization) was seen to be their favorite. Indeed, the variational Bayesian inference is one of the widely used techniques in parameters estimations. It is a deterministic approach that guarantees convergence. So, the method is efficient, but it is not very accurate due to the strong independency assumption (between latent variables and parameters) often observed in the variational Bayes methods. It usually leads to the traditional factorization or the decoupling of the joint variational distribution into a product of individual variational distributions. So, when there is dependency between latent variables and parameters, the variational Bayes becomes inaccurate as it could severely affect the lower bound and jeopardize estimation when this lower bound become instable, affecting the log likelihood computation. A solution proposed by Teh et al. and Caballero et al. in [12, 18] was to marginalize out the parameters leaving only the latent variables that could be now assumed independent given these parameters. So these authors provided a weak assumption which is more robust for exact inference. It leads to the collapsed variational inference where the parameters are marginalized out. The only drawback with the inference was still the Dirichlet distribution.

The CVB-LGDA we finally proposed in this chapter has a graphical architecture that seems to be similar to the bayesian hierarchical model proposed by Fei-Fei Li et al. in 2005 in [2]. However, there is a major difference between these two classification models. In fact, the model proposed in [2] draws its document parameters from the Dirichlet distribution while our new model samples its corpus and documents parameters from the GD.

As a result, with the GD, we automatically improve the previous state-of-the-art inference which was a Dirichlet-based inference. The new collapsed variational Bayesian inference in this chapter is now a generalized Dirichlet-based one. It is more robust and versatile for a better topic correlation and codeword's analysis as this will help in the intra class variation problem. The experimental results have shown the new model performs better than the one proposed by [2]. Though, credits could be highly given to these authors in [2, 3] for their tremendous effort to provide an early exploration of the LDA model that ultimately led to these several LDA extensions observed today. Their model and algorithm could be summarized through these following concepts: For instance, given observed variables u and unobserved or latent variables x and the model parameters θ we are maximizing the loglikelihood with respect to θ such that:

$$\mathscr{L}(\theta) = \log p(u|\theta) = \log \int p(x, u|\theta) dx \tag{3}$$

Often, the difference between the loglikelihood and the bound can be expressed as:

$$\begin{aligned}\mathscr{L}(\theta) - \mathcal{F}(\tilde{Q}(x), \theta) &= \log p(u|\theta) - \int \tilde{Q}(x) \log \frac{p(x,u|\theta)}{\tilde{Q}(x)} dx && (4)\\ &= \log p(u|\theta) - \int \tilde{Q}(x) \log \frac{p(x|u,\theta)p(u|\theta)}{\tilde{Q}(x)} dx && (5)\\ &= - \int \log \frac{p(x|u,\theta)}{\tilde{Q}(x)} dx && (6)\\ &= KL(\tilde{Q}(x), p(x|u,\theta)) && (7)\end{aligned}$$

This difference is actually the Kullback-Leibler divergence. It is non negative and zero if and only if $\tilde{Q}(x) = p(x|y,\theta)$ (this is the E-step). Based on the bound on the log-likelihood; this likelihood is non decreasing in every iteration such that:

$$\mathscr{L}(\theta^{(k-1)}) \underbrace{=}_{E-step} \mathcal{F}(Q^{(k)}, \theta^{(k-1)}) \underbrace{\leq}_{M-step} \mathcal{F}(\tilde{Q}^{(k)}, \theta^{(k)}) \underbrace{\leq}_{Jensen \quad inequality} \mathscr{L}(\theta^{(k)}) \qquad (8)$$

where EM converges to a local optimum of $\mathscr{L}$. The variational Bayes EM implemented in Algorithm.2 while ours (MCMC) is illustrated by Algorithm.4 mainly shows the difference in the two models.

Algorithm 2 Variational Bayes Expectation-Maximization (EM)

```
Goal: lower bound $p(u|m)$
$\boldsymbol{VB-E \ \ step}$: compute the variational parameters such that
$\tilde{Q}_x^{(t+1)}(x) = p(x|u,\theta^{(t)})$
$\boldsymbol{VB-M \ \ step}$: compute the parameters using the variational estimates from E-step as:
$\tilde{Q}_{(\theta)}^{(t+1)}(\theta) \propto \exp(\int \tilde{Q}_{(x)}^{(t+1)}(x) \log p(x,u,\theta) dx$
```

Therefore, although using similar graphical topic model for classification where the vocabulary is shared among all classes, the priors and the inferences are different using Fei-Fei Li et al.([2]) and our method. In other words, the models are different.

Algorithm 3 summary of the CVB-LGDA Inference

```
procedure
Input: $\mathcal{X}$, $\varepsilon = (\alpha_c, \beta_c)$, $iterMax$, $\zeta = (\lambda, \eta)$, $K$, $V$, $N$
Initialize $Z$, $N_{jk.}$, $N_{.kv_{ij}}$
    for $iter$ = 1 to $iterMax$ do
        for $i = 1$ to $N$ in document $j$ in class $c$ do
            $z_{ij} \sim \hat{Q}(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ using Eq.54
            Update $N_{kv}^t$, $N_k^t$, $N_{dk}^t$
        end for
    end for
Output: Parameters $\tilde{\theta}_{jks}$ and $\tilde{\varphi}_{kws}$ using Eq.55 and 56
end procedure
```

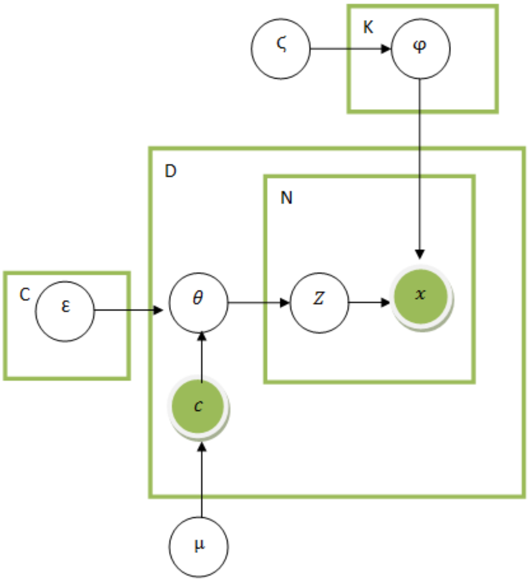


Figure 2.1: Topic Graphical Model for Classification. The shaded circle denotes observed variables $\boldsymbol{x}$ and the class c.

2.3 The New Approach

2.3.1 Overview

In this chapter, due to the limitations of the Dirichlet prior, we propose the generalized Dirichlet (GD) distribution on both the document and corpus parameters for its flexibility [23, 36] in a collapsed space: the GD has a more general and versatile covariance structure than the Dirichlet prior. In addition, the Dirichlet is a special case of the GD. A variational inference scheme with this conjugate prior in the collapsed space represents an improvement to the state-of-the-art in images, 3D objects, and videos analysis to deal with challenges related to extensive vocabulary size, and increasing number of topics. The new approach integrates two models: a topic model (unsupervised learning) and a classification model (supervised learning). The topic graphical model (Fig. 2.1) in this classification problem is described by a list of variables as shown below. It shows the conditional dependence structure between these variables. Moreover, as we are planning to implement inferences in these two following spaces, details about the collapsed and the joint spaces will be provided in this section. Meanwhile, back to our graphical model that is a directed acyclic graph, the variables are indeed described as follows:
D-Number of documents
N-Number of words in each document
K-Number of topics
$\boldsymbol{x} = \{x_{ij}\}$-Observed words (where a word is positioned as ith in the jth document)
$z = \{z_{ij}\}$-latent variables (topic indices) associated to the observed words $\{x_{ij}\}$
$\theta_j = \{\theta_{jk}\}$-Mixing proportions (each parameter θ_j is a mixture of K topics)
$\varphi_k = \{\varphi_{kw}\}$- Corpus parameters
$\theta_{jk}/\varepsilon \sim GenDir(\varepsilon)$-Generalized Dirichlet distribution with hyperparameter ε for the

document parameter θ_{jk}
$\varphi_{kw}/\zeta \sim GenDir(\zeta)$-Generalized Dirichlet distribution with hyperparameter ζ for the corpus parameter φ_{kw}
$z_{jk}/\theta_{jk} \sim Mult(\theta_{jk})$-Multinomial distribution with parameter (θ_{jk})
$x_{jk}/z_{jk}, \varphi_{jk} \sim Mult(\varphi_{kw})$-Multinomial distribution with parameter φ_{kw}
$\boldsymbol{c} = \{1, 2, ..., C\}$ is the set of all classes summarizing the database, similar to [2].
$(\varepsilon, c) = (\alpha_{c1}, \beta_{c1}, ..., \alpha_{cK}, \beta_{cK}) = (\alpha_c, \beta_c)$
$\zeta = (\lambda_1, \eta_1, ..., \lambda_V, \eta_V) = (\lambda, \eta)$

In this chapter, the documents are drawn from a class set c. The variables ε and ζ are the documents and corpus hyperparameters of the graphical model, respectively, using the generalized Dirichlet as priors. In implementation, the variable ε holds two $C \times K$ matrices α and β such that ε_c is K-dimensional GD hyperparameter (α_c, β_c) for the document. Similarly, for a every topic k, the variable ζ contains two vectors of size $V \times 1$, (λ and η) such that ζ is a V-dimensional GD hyperparameter (λ, η) for the corpus using the vocabulary of size V. In addition, the CVB-LGDA algorithm uses notions of variational distributions and variational lower bound. In our new scheme and similar to [30], the variable $\tilde{Q}$ is the variational distribution in the standard space (the joint space of parameters and latent variables). However, the distribution $\hat{Q}$ is the variational in the collapsed space of latent variables where the parameters are marginalized out. In the exponential family distribution, typical to many LDA related graphical model distributions, the marginal likelihood function (the normalization factor in the posterior distribution) is often approximated by a lower bound defined as $\exp(\mathcal{F}(Q(x)))$, where $\mathcal{F}(Q(x))$ is the variational lower bound in the log space [37]. This element of the integration functional is also called the variational free energy [30, 12]. Our model is an improved variational Bayes approach in the collapsed space of latent variables. The traditional VB inference is performed in the joint space of latent variables and model parameters. Though, it is slow compared to the VB in the collapsed space. We therefore define these bounds to clarify all the steps taken for the implementation of the new approach in the collapsed space (in comparison to the joint space). As a result, similar to $\tilde{Q}$ and $\hat{Q}$, the variable $\tilde{\mathcal{F}}$ is the variational bound in the joint space while $\hat{\mathcal{F}}$ is the variational bound in the collapsed space using our CVB-LGDA graphical model (Fig. 2.1). This concept is similar to [30].

2.3.1.1 Notations and definitions

In this classification problem, it is important to define some basic concepts related to the BoW framework as we deal with different data types such as images, 3D objects, and videos. A video sequence can be seen as a collection of frames (images). Since an image and a 3D object are each assimilated to a document, a *patch* x is defined as the basic unit for a document; and it is an element of the vocabulary codewords. The document is reduced to a sequence of vocabulary codewords (after quantization scheme from the clustering algorithm). Therefore, an *image, 3D object or a video frame* $\mathcal{X}$ are each a collection of N patches defined as $\mathcal{X} = (x_1, x_2, ..., x_N)$. The variable x_n is the n^{th} patch in the image. A *category or a class* is a collection of D images such that $I = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_D\}$. In our image, 3D object, and video analysis within the BoW, the document is a collection of patches. Though, in 3D object analysis, the document is also defined as a sequence of images or 2D views which in turn are a collection of patches. Therefore, our model from image analysis could be easily generalized to a 3D object and a video as we treat each 3D

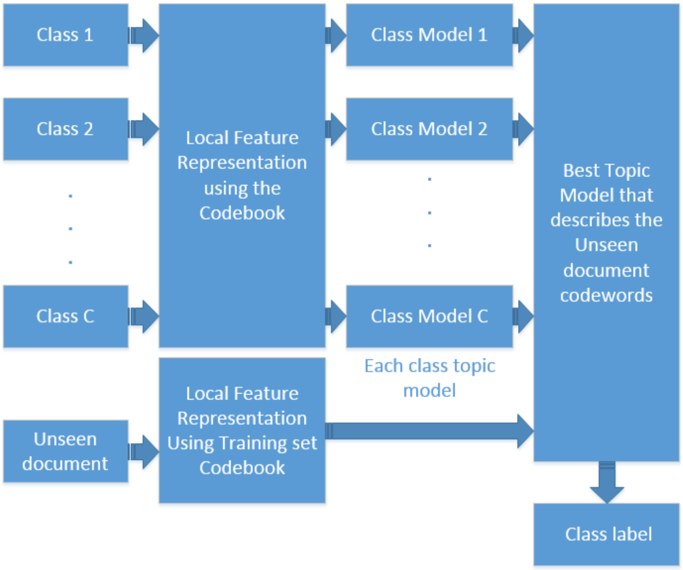

Class 1
Class 2
Class C
Local Feature Representation using the Codebook
Class Model 1
Class Model 2
Class Model C
Each class topic model
Best Topic Model that describes the Unseen document codewords
Unseen document
Local Feature Representation Using Training set Codebook
Class label

Figure 2.2: Topic model for classification problem

or video documents as a sequence of 2D views within the BoW structure.

2.3.2 Proposed topic model

In this chapter, our GD-based collapsed variational Bayesian approach utilizes a topic modeling scheme for a classification problem using images, 3D objects and videos. Most importantly, this classification approach emphasizes on the generative probabilistic model as it has ability to learn both the class-conditional probability $p(\mathcal{X}|c, \varepsilon, \zeta)$ and the prior probability $p(c|\mu)$ (Eq. 113) before estimating the posterior distribution $p(c|\mathcal{X}, \varepsilon, \zeta, \mu)$ using the Bayes' rule. It is for instance in contrast to the discriminative model that usually learns directly the posterior distribution $p(c|\mathcal{X}, \varepsilon, \zeta, \mu)$ [38]. As a result, in our new framework, each class-conditional probability is a topic model that learns its codewords distribution. With the GD conjugate prior to the multinomial, the new method aims to capture semantic relationships between vocabulary words and between topics. So, through this effective representation, the model could easily be generalized to several other applications. Again, as a contribution, this chapter extends the capabilities of the previous CVB technique by introducing a better prior that facilitates applications using images, 3D objects, and videos. The topic modeling scheme (GD-based CVB) in this categorization problem ultimately provides the best model describing codewords distribution of the observed data in each class. In this section, we will also present the GD distribution and its advantages over the Dirichlet prior.

In computer science, the time complexity is the computational complexity that describes the amount of time it takes to run an algorithm. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, supposing that each elementary operation takes a fixed amount of time to perform. Thus, the amount of time taken and the number of elementary operations performed by the algorithm are taken to differ by at most a constant factor.

When dealing with large scale collection of datasets and their processing, two notions that often come in mind are the time complexity and the memory or space complexity. The time complexity coul be defined as the computational complexity describing the amount of time it takes to execute or run an algorithm. The concept ultimately counts the number of elementary operations performed by the algorithm. The memory complexity is the memory used by an algorithm. In this chapter, we are deeply interested in the time complexity in our model as we observe its performance with real data in comparaison to the other classification models. In a topic model with K as total number of topics, and N the humber of unique words in the dataset, and V as vocabulary size, we can observe that the LDA and the CVB-LDA have similar time complexiy noted as $O(NK)$ implying these models have ability to generate topics. Importantly, the GD-LDA also has the same time complexity $O(KN)$. While the CVB-LDA and the LDA could only generate topics, the GD-LDA could with the same amount of time, perform two tasks: semantic relationship between codewords, and topic correlation analysis. Same time complexity is observed by the LGDA model. In our model, the CVB-LGDA emphasizes on topic correlation, semantic relationship between words, and codebook analysis bringing his overall time complexity to $O(NKV)$ as seen in Eq. 9. Though the flexibility of the CVB-LGDA allows it to prune out irrelevant topics and irrelevant vocabulary codewords reducing then the vocabulary size. Therefore, as K and V can be extremely small due to pruning, $O(NKV)$ could be reduced to $O(N)$ Eq. 10. The new mdel can get more done faster.

In addition, the variational Bayes-based methods despite its efficiency could be very slow as the inference operates in the joint space of the latent variables and the parameters whereas the new approach operating in the collapsed space is gets its parameters marginalized out leaving only the latent variables. We finally conclude the new approach has potential to be faster than its competitors as it still takes advantage of the Taylor approximation to speed up computation. The models along with their time complexity summarized in Table 2.2

$$\left.\begin{aligned} for\ n \quad &= 1:N \\ for\ k \quad &= 1:K \\ for\ v \quad &= 1:V \end{aligned}\right\} \rightarrow timeComplexity = O(NKV) \tag{9}$$

The objective of this model is that for very large value N (data size) and very small value of K and V we can reach a linear time $O(N)$ therefore for

$$\left.\begin{aligned} K \quad &\ll N \\ V \quad &\ll N \end{aligned}\right\} \rightarrow timeComplexity = O(N) \tag{10}$$

2.3.3 Inference schemes

This section is dedicated to the inference techniques in the new method. In addition, it includes the different inference schemes used in the previous extensions.

2.3.3.1 General Bayesian inference procedures with VB and CGS

The goal in any Bayesian framework is the computation of the posterior distribution in inferences. However, and very often, it involves integrals estimations such as the likelihood function and the model posterior distribution that are not quite tractable. Therefore, several schemes such as VB with EM algorithm and MCMC are widely used to uncover the topics and estimate the model parameters. Each of these methods has its advantages,

	time complexity	Analysis
LDA	O(NK)	no topic correlation
CVB-LDA	O(NK)	no topic correlation
GD-LDA	O(NK)	topic correlation leading to O(N)
LGDA	O(NK)	topic correlation leading to O(N)
CTM and PAM	$O(K^2N)$	topic correlation but very expensive $O(N)$
CVB-LGDA	O(NKV)	topic correlation and vocabulary analysis leading to O(N) as time complexity when the number of topics and vocabulary size are reduced
KNN	O(KNDim)	No topic correlation as K refers to the K-nearest neighbors (not topics), and Dim is the data dimensionality
SVM	$O(N^3)$	topic correlation but very expensive $O(N^3)$
BPNN	$O(N^5)$	topic correlation but very expensive $O(N^5)$

Table 2.2: Time complexity between the new CVB-LGDA model and the other schemes within the BoW framework

but also its drawbacks. The state-of-the-art seems to reconcile the advantages of both VB and the Gibbs sampler in the collapsed space, leading to an hybrid model which represents the best of both worlds: the collapsed Variational Bayes (CVB) inference. It is intuitively a variational Bayes approach in the collapsed space of latent variables using the Gibbs sampler. The CVB inference ultimately solves the problem of convergence in the MCMC approach. In addition, it removes the bias in the VB method with an inference scheme in exact fashion where the latent variables are conditionally independent given the parameters [30]. From the graphical model in Fig. 2.1, given its hyperparameters ε, ζ, and the class parameter μ, we can express the full generative equation of the model. It is the joint probability distribution noted $p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu)$ and illustrated below as:

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) \prod_{i=1}^{K} p(\varphi_i|\zeta) \prod_{j=1}^{D} p(\theta_j|\varepsilon, c) \times \prod_{n=1}^{N} p(z_{j,n}|\theta_j) p(x_{j,n}|\varphi z_{j,n}) \qquad (11)$$

This joint distribution's equation can be simplified to :

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) p(\theta|c, \varepsilon) p(\varphi|\zeta) \times \prod_{n=1}^{N} p(z_n|\theta) p(x_n|z_n, \varphi) \qquad (12)$$

where $p(\varphi|\zeta)$ and $p(\theta|c, \varepsilon)$ are the corpus prior distribution (GD) with hyperparameters ζ and a class document prior distribution (GD) with hyperparameter ε, respectively. The distributions $p(z_n|\theta)$ and $p(x_n|\varphi z_n)$ are multinomial while the distribution $p(c|\mu)$ is the class prior. The Bayesian inference approximates the posterior distribution of the latent variables z and the model parameters θ and φ given the observations and the class. This is the joint posterior distribution $p(z,\theta,\varphi|\mathcal{X},c,\varepsilon,\zeta,\mu)$ as shown in the equation below.

$$p(z,\theta,\varphi|\mathcal{X},c,\varepsilon,\zeta,\mu) = \frac{p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta,\mu)}{p(\mathcal{X},c|\varepsilon,\zeta,\mu)} \tag{13}$$

where the denominator is expressed as :

$$p(\mathcal{X},c|\varepsilon,\zeta) = \int_\theta \int_\varphi \sum_z p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta) d\varphi d\theta \tag{14}$$

with

$$p(\mathcal{X},c|\varepsilon,\zeta,\mu) = p(\mathcal{X}|\varepsilon,\zeta,c)p(c|\mu) \tag{15}$$

For a uniform class prior, we obtain $p(c|\mu) = p(c) = \frac{1}{C}$ with μ negligible. As a result, the Eqs. 71 and 72 could be simplified as :

$$p(\mathcal{X},c|\varepsilon,\zeta,\mu) = \frac{p(\mathcal{X}|\varepsilon,\zeta,c)}{C} \tag{16}$$

C is the total number of classes while c is the set of classes in this graphical model. The posterior distribution is then reduced to :

$$p(z,\theta,\varphi|\mathcal{X},c,\varepsilon,\zeta,\mu) = \frac{p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta,\mu)}{p(\mathcal{X}|\varepsilon,\zeta,c)/C} \tag{17}$$

As the likelihood function here, the class conditional $p(\mathcal{X}|c,\varepsilon,\zeta)$ is not tractable, the posterior $p(z,\theta,\varphi|\mathcal{X},c,\varepsilon,\zeta,\mu)$ is not tractable as well. Then, the variational Bayes (VB) estimates the true posterior distribution using variational distributions [30] (factorized distributions) $\tilde{Q}(z,\theta,\varphi)$ such that:

$$\tilde{Q}(z,\theta,\varphi) = \prod_{ij} \tilde{Q}(z_{ij}|\tilde{\psi_{ij}}) \prod_j \tilde{Q}(\theta_j|\tilde{\varepsilon_j}) \prod_k \tilde{Q}(\varphi_k|\tilde{\zeta_k}) \tag{18}$$

where $\tilde{Q}(z_{ij}|\tilde{\psi_{ij}})$ is the variational multinomial distribution with parameters $\tilde{\psi_{ij}}$. However, $\tilde{Q}(\theta_j|\tilde{\varepsilon_j})$ and $\tilde{Q}(\varphi_k|\tilde{\zeta_k})$ are the GD variational distributions with parameters $\tilde{\varepsilon_j}$ and $\tilde{\zeta_k}$, respectively, in the joint space of latent variables and model parameters. This VB was often implemented in LDA and LGDA.

As the standard VB operates in the joint space of latent variables and parameters, inference in that space requires a family of distributions, a set of variational distributions, defined as $\tilde{Q}(z,\theta,\varphi)$ that are as close as possible or tight to the true posterior distribution $p(z,\theta,\varphi|c,\varepsilon,\zeta)$ with the KL (KullBack Leibler) divergence. Importantly, VB introduces a lower bound to the marginal log likelihood, a concept that is also equivalent to the VB upper bounding the negative log marginal likelihood $-\log p(\mathcal{X}|c,\varepsilon,\zeta)$ in a framework [30] that utilizes variational free energy as shown in Eqs. 77 and 80. The inference leads to variational parameters updates and the model parameters estimation. VB is efficient as it is easy to implement and provides an easy access to convergence. It is a deterministic approach. From Eqs. 77 to 81, the bound on the loglikelihood is expressed as:

$$\begin{aligned}\log p(\mathcal{X}|c,\varepsilon,\zeta) \geq & \int_\theta \int_\varphi \sum_z Q(z,\theta,\varphi) \log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta) d\varphi d\theta \\ & - \int_\theta \int_\varphi \sum_z Q(z,\theta,\varphi) \log Q(z,\theta,\varphi) d\varphi d\theta \\ & = \mathbb{E}_Q[\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - \mathbb{E}_Q[\log Q(z,\theta,\varphi)]\end{aligned} \tag{19}$$

$$\begin{aligned}-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq & -\int_\theta \int_\varphi \sum_z Q(z,\theta,\varphi) \log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta) d\varphi d\theta \\ & + \int_\theta \int_\varphi \sum_z Q(z,\theta,\varphi) \log Q(z,\theta,\varphi) d\varphi d\theta \\ & = \mathbb{E}_Q[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - \mathbb{E}_Q[-\log Q(z,\theta,\varphi)]\end{aligned} \tag{20}$$

$$-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq \tilde{\mathcal{F}}(\tilde{Q}(z,\theta,\varphi)) = \mathbb{E}_{\tilde{Q}}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - \mathcal{H}(\tilde{Q}(z,\theta,\varphi)) \tag{21}$$

As the variational entropy is expressed as $\mathcal{H}(\tilde{Q}(z,\theta,\varphi)) = \mathbb{E}_{\tilde{Q}}[-\log \tilde{Q}(z,\theta,\varphi)]$, the variational posterior distribution in the joint space $\tilde{Q}(z,\theta,\varphi)$ is factorized using the independency assumption as shown in Eq. 75. Consequently, in the joint space of VB using a GD prior, estimating the model parameters θ, φ (in M step) from a variational EM algorithm requires approximation and update of the GD variational distributions hyperparameters when using the variational multinomial parameter $\tilde{\psi}_{ijkc}$ in the E-step. In terms of inferences, many researchers have implemented the Dirichlet-based VB [3, 2, 30, 39, 23], but its limitations (strong independency assumption) ultimately led to the Dirichlet-based CVB which is a combination of VB and MCMC approaches.

In general, the CVB [30, 40, 13] is an improved version of the VB in the collapsed space of latent variables; and it is the state-of-the-art inference we are also upgrading because of the limitation of its Dirichlet prior. The CVB and the CGS both operate in the collapsed space. Therefore, from the joint distribution $p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta,\mu)$, the model parameters θ, φ are integrated out to obtain the marginal distribution $p(\mathcal{X},z,c|\varepsilon,\zeta)$ defined as :

$$p(\mathcal{X},z,c|\varepsilon,\zeta) = \int_\theta \int_\varphi p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta) d\varphi d\theta \tag{22}$$

But $p(\mathcal{X},z,c|\varepsilon,\zeta) = p(\mathcal{X},z|c,\varepsilon,\zeta)p(c)$ so $p(\mathcal{X},z|c,\varepsilon,\zeta)$ becomes

$$p(\mathcal{X},z|c,\varepsilon,\zeta) = C \int_\theta \int_\varphi p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta) d\varphi d\theta \tag{23}$$

Due to the prior conjugacy between the GD and the multinomial distributions, this integral is easy to compute, and is often expressed as a product of gamma functions. The goal is to approximate the conditional distribution of the latent variable $p(z|\mathcal{X},c,\varepsilon,\zeta)$.

2.3.3.2 The New Collapsed Gibbs sampler and Mean field inference

The collapsed space of latent variables is a low dimensional space. The space is suitable for easy computation of integrals using the conjugacy property between the priors distributions and the multinomial distributions. Ultimately, the Gibbs sampler provides inference by computing expectations through a sampling process of the latent variables to approximate the posterior distributions using a network of conditional probabilities (Bayesian network). The CGS [30, 18, 41] in the collapsed space of latent variables is therefore very fast compared to the standard Gibbs in the joint space of latent variables and model parameters. In addition, with the CGS, no more use of digamma functions which were computationally very expensive in VB method. The CGS algorithm estimates the parameters when the Markov chain reaches its stationary state (stationary distribution) and provides the best estimate of the true posterior distribution.

From the marginal joint distribution $p(\mathcal{X}, z|c, \varepsilon, \zeta)$, the conditional probabilities of the latent variable z_{ij} are computed given the current state of all variables except the particular variable z_{ij} being sampled [30]. The scheme uses the collapsed Gibbs sampler for topic assignments. The conditional probability of latent variables is $p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ where $-ij$ corresponds to counts or variables with z_{ij} excluded [30]. This conditional probability is expressed as :

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) = \frac{p(z_{ij}, z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)}{p(z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)} \tag{24}$$

The above equation using [30] can be simplified since:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) \propto p(z_{ij} = k, z^{-ij}, \mathcal{X}, c|\varepsilon, \zeta) \tag{25}$$

The obtained Callen equations (below) as in [30] illustrate the way the collapsed Gibbs actually performs the sampling mechanism. It is an expectation problem as shown in the equation given as:

$$p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta) = \mathbb{E}_{p(z^{-ij}|c,\mathcal{X},\varepsilon,\zeta)}[p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)] \tag{26}$$

2.3.3.3 Using GD in the collapsed Gibbs sampler

In our model, the parameters θ, φ are drawn from the generalized Dirichlet distribution. These parameters are now marginalized out in the collapsed space of the latent variables to speed up sampling process. It is faster to sample in the collapsed space than in the joint space of latent variables and parameters [30]. The motivation here is to sample the latent variables from the joint distribution $p(\mathcal{X}, z|c, \varepsilon, \zeta)$ using a network of single class conditional probabilities illustrated below. As previously mentioned, the conjugacy assumption facilitates estimation of this integral obtained as a product of gamma functions (Eq. 127).

$$\begin{aligned} p(\mathcal{X}, z|c, \varepsilon, \zeta) = C \prod_{j=1}^{D} \left[\prod_{i=1}^{K} \frac{\Gamma(\alpha_{ci} + \beta_{ci})}{\Gamma(\alpha_{ci})\Gamma(\beta_{ci})} \prod_{i=1}^{K} \frac{\Gamma(\alpha'_{ci})\Gamma(\beta'_{ci})}{\Gamma(\alpha'_{ci} + \beta'_{ci})} \right] \\ \times \prod_{j=1}^{D} \left[\prod_{i=1}^{K} \frac{\Gamma(\lambda_r + \eta_r)}{\Gamma(\lambda_r)\Gamma(\eta_r)} \prod_{i=1}^{K} \frac{\Gamma(\lambda'_r)\Gamma(\eta'_r)}{\Gamma(\lambda'_r + \eta'_r)} \right] \end{aligned} \tag{27}$$

where the document-topic update in class c is expressed as :

$$\begin{cases} \alpha'_{ci} = \alpha_{ci} + N^{i}_{j(.)} \\ \beta'_{ci} = \beta_{ci} + \sum_{l=i+1}^{K+1} N^{l}_{j(.)} \end{cases} \tag{28}$$

The topic-word update is defined as :

$$\begin{cases} \lambda'_r = \lambda_r + N^i_{(.),r} \\ \eta'_r = \eta_r + \sum_{d=v+1}^{V+1} N^i_{(.)d} \end{cases} \tag{29}$$

These update equations above are observed to be very similar to the updates expected from the variational inference. However, the current multinomial updates are provided by the Gibbs sampler (Eqs. 128, 142, and 30)

$$\begin{cases} N^i_{j(.)} = N^{ij}_{jk(.)} = N^{ij}_{jk.} \\ N^l_{j(.)} = N^{ij}_{jl(.)} = N^{ij}_{jl.} \\ N^i_{(.),r} = N^{ij}_{(.),k\nu_{ij}} = N^{ij}_{.k\nu_{ij}} \\ N^i_{(.)d} = N^{ij}_{(.),kd} = N^{ij}_{.kd} \end{cases} \tag{30}$$

where i refers to the i^{th} topic in document j . The variable l indexes $(k+1)^{th}$ topic in document j. The variable r refers to the v^{th} codeword in topic k while d refers to the $(v+1)^{th}$ codeword in topic k. The count $N^{ij}_{jk.}$ is the number of word i in the document j in topic k in class c. In addition, $N^{-ij}_{jk.}$ is the total number of words in topic k in document j in class c except the word i being sampled. The constant $N^{ij}_{.k\nu_{ij}}$ is the number of times the codeword ν appears in topic k in document j while $N^{-ij}_{.k\nu_{ij}}$ is the number of times the word ν appears in document j in topic k except the one being sampled.

In Eq. 96, we obtained the sampling equation of a topic z^{ij} in a particular class document j given the observations x and the initial topic assignments associated to each word except the one being sampled z^{-ij}. The counts in the document-topic and topic-word structure are ultimately emphasized by the multinomial variable $\hat{\psi}_{ijk}$ in the Gibbs sampler, similar to the case of the VB. Though, the count in Eq. 96 is obtained in a collapsed space, so it is different from the one in the joint space of VB. As parameters are marginalized out in a particular class, the update is reduced to:

$$\hat{\psi}_{ijkc} = p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta) \tag{31}$$

using $p(z_{ij}|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) = \frac{p(z_{ij}, z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)}{p(z^{-ij}, \mathcal{X}, c|\varepsilon, \zeta)}$ from Eq. 90 so that:

$$\begin{aligned} p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) \propto & \left[\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N^{-ij}_{jl.})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N^{-ij}_{jl.})}\right] \\ & \times \left[\frac{(N^{-ij}_{.k\nu_{ij}} + \lambda_\nu)(\eta_\nu + \sum_{d=\nu+1}^{V+1} N^{-ij}_{.kd_{ij}})}{(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} N^{-ij}_{.kd_{ij}})}\right] \end{aligned} \tag{32}$$

Normalizing the distribution above leads to a posterior probability defined as :

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta) = \frac{A(k)}{B(k', K)} \tag{33}$$

such that :

$$A(k) = \left[\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N^{-ij}_{jl.})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N^{-ij}_{jl.})} \times \frac{(N^{-ij}_{.k\nu_{ij}} + \lambda_\nu)(\eta_\nu + \sum_{d=\nu+1}^{V+1} N^{-ij}_{.kd_{ij}})}{(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} N^{-ij}_{.kd_{ij}})}\right] \tag{34}$$

and

$$B(k', K) = \sum_{k'=1}^{K} \left[\frac{(N_{jk'.}^{-ij} + \alpha_{ck'})(\beta_{ck'} + \sum_{l=k'+1}^{K+1} N_{jl.}^{-ij})}{(\alpha_{ck'} + \beta_{ck'} + \sum_{l=k'}^{K+1} N_{jl.}^{-ij})} \; \frac{(N_{.k'\nu_{ij}}^{-ij} + \lambda_{\nu})(\eta_{\nu} + \sum_{d=\nu+1}^{V+1} N_{.k'd_{ij}}^{-ij})}{(\lambda_{\nu} + \eta_{\nu} + \sum_{d=\nu}^{V+1} N_{.k'd_{ij}}^{-ij})} \right] \tag{35}$$

Now, the collapsed Gibbs sampler uses the Callen equations (Eq. 92) as in [30] to sample z given the observable variable $\mathcal{X}$. This equation implies that the conditional $p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta)$ are approximated through sample mean of $p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ by drawing enough $p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ such that the variables z^{-ij} are in turn drawn from probability distribution $p(z^{-ij}|\mathcal{X}, c, \varepsilon, \zeta)$. In other words, it is the expected value of $p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ where samples are drawn from $p(z^{-ij}|\mathcal{X}, c, \varepsilon, \zeta)$. The Gibbs sampling is equivalent to an approximation of the true posterior distribution (in a Bayesian inference) in the collapsed space. As a result, in the CGS, the expected multinomial parameter in each class is estimated as a count from the true posterior distribution in the Eq. 96. As the CGS samples from the true posterior distribution in the collapsed space, the VB updates its variational parameters in the joint space of the latent variables and model parameters using the expected multinomial parameter $\tilde{\psi}_{ijkc}$. Therefore,

$$\tilde{\psi}_{ijkc} \neq \hat{\psi}_{ijkc} \tag{36}$$

2.3.3.4 The GD-based variational Bayes: GD-VB

As a deterministic approach and in contrast to the CGS, the VB insures convergence to a local minimum. Optimizing the variational distribution in Eq. 75 from Eq. 81 with respect to the GD variational parameters leads to the following updates in the parameters of the corpus and documents GD variational distributions. These updates are similar to the CVB-LDA [30].

$$\tilde{\alpha}_{jkc} = \alpha_c + \sum_i \tilde{\psi}_{ijkc} \tag{37}$$

$$\tilde{\beta}_{jk'c} = \beta_c + \sum_i \tilde{\psi}_{ijk'c} \tag{38}$$

$$\tilde{\lambda}_{kw} = \lambda + \sum_{ij} \vec{1}(x_{ij} = w)\tilde{\psi}_{ijkc} \tag{39}$$

$$\tilde{\eta}_{kw'} = \eta + \sum_{ij} \vec{1}(x_{ij} = w')\tilde{\psi}_{ijkc} \tag{40}$$

where $k' = k + 1$ and w' are respectively the $(k+1)^{th}$ topic in the document and the $(v+1)^{th}$ codeword in the vocabulary. The multinomial update (count) $\tilde{\psi}_{ijkc}$ is also obtained through optimization of the joint posterior variational distribution $\tilde{\mathcal{F}}(\tilde{Q}(z))$ with respect to the multinomial variational parameter [30].

In the joint space, the document GD variational parameter $\tilde{\alpha}_{jkc}$ is a document-topic count; it is the total number of words in a topic k in a document j, all in a class c. The GD variational parameter $\tilde{\beta}_{jkc}$ is also a document-topic count. It is the total number of words from the next $(k+1)^{th}$ topic up to the total number of topics in a document j in class c. The corpus GD variational parameter $\tilde{\lambda}_{kw}$ is a word-topic count: it is the number of times a word w (a codeword from a vocabulary of size V) appears in the topic k in a document j. Similarly, $\tilde{\eta}_{kw'}$ is another word-topic count as it is the total number of words left in the vocabulary once the $(v+1)^{th}$ word is selected such that the first v words are not counted. These variational parameters are updated with the variational multinomial parameter $\tilde{\psi}_{ijkc}$. Despite its efficiency with a well defined convergence criterion [30, 3, 5], the VB often suffers for large bias (strong independency assumption) as it decouples the joint variational posterior into a product of individual variational posterior distributions. This is because the model always neglects (for convenience) to consider that the latent variables and model parameters could be dependent in the true posterior distribution. The situation could make inferences (posterior distribution estimation) inaccurate as the lower bound in this case is no longer robust. In addition, the VB is not always capable of implementing a proper mean field approximation (inference), because the scheme ultimately operates in the joint space of latent variables and parameters such that any change in the parameters could affect the latent variables [30]. Considering efficiency and accuracy, the new technique, the GD-CVB combines the advantages of both GD-VB and GD-CGS. The approach operates in the collapsed space of the latent variables.

2.3.3.5 The new GD-based Collapsed variational Bayes (CVB) architecture: Mean field variational inference

It is a GD-based VB in the collapsed space (GD-CVB inference). This new collapsed variational Bayes inference (of the CVB-LGDA model) is a combination of the GD-based VB and GD-based CGS. Similar to [30], the GD-CVB inference procedure models the dependence of parameters related to the latent variables in an exact fashion where parameters are either marginalized out in the graphical representation or modeled as the joint $p(\theta, \varphi | z, \mathcal{X}, c, \varepsilon, \zeta)$. It leaves the latent variables weakly dependent, therefore assumed independent. As a result, through this weak assumption, the GD-CVB provides an efficient framework for mean field approximation as latent variables are conditionally independent given the parameters. Then, based on the conditionally independence assumption of the latent variables, a better set of variational distributions could be obtained as this weaker assumption allows to finally decouple effectively the joint $\hat{Q}(z, \theta, \phi)$. It is given as :

$$\hat{Q}(z, \theta, \varphi) = \hat{Q}(\theta, \varphi | z) \prod_{ij} \hat{Q}(z_{ij} | \hat{\psi}_{ij}) \tag{41}$$

where $\hat{Q}(z_{ij} | \hat{\psi}_{ij})$ is the variational multinomial distribution with parameters $\hat{\psi}_{ij}$ in the collapsed space, and the variational free energy $\hat{\mathcal{F}}(\hat{Q}(z)\hat{Q}(\theta, \varphi | z))$ conditional to z becomes:

$$\hat{\mathcal{F}}(\hat{Q}(z)\hat{Q}(\theta, \varphi | z)) = \mathbb{E}_{\hat{Q}(z)\hat{Q}(\theta, \varphi | z)}[-\log p(\mathcal{X}, z, \theta, \varphi, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{Q}(z)\hat{Q}(\theta, \varphi | z)) \tag{42}$$

$$\hat{\mathcal{F}}(\hat{Q}(z)\hat{Q}(\theta, \varphi | z)) = \mathbb{E}_{\hat{Q}(z)}[\mathbb{E}_{\hat{Q}(\theta, \varphi | z)}[-\log p(\mathcal{X}, z, \theta, \varphi, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{Q}(\theta, \varphi | z))] - \mathcal{H}(\hat{Q}(z)) \tag{43}$$

With only two variational posterior distributions ($\hat{Q}(\theta, \varphi | z)$, and $\hat{Q}(z)$), the variational free energy is minimized with respect to $\hat{Q}(\theta, \varphi | z)$ and then with respect to the collapsed variational $\hat{Q}(z)$ as shown in [30]. A minimum variational free energy is reached at the true posterior $\hat{Q}(\theta, \varphi | z) = p(\theta, \varphi | z, \mathcal{X}, c, \varepsilon, \zeta)$ which becomes :

$$\hat{\mathcal{F}}(\hat{Q}(z)) \triangleq \min_{\hat{Q}(\theta, \varphi | z)} \hat{\mathcal{F}}(\hat{Q}(z)\hat{Q}(\theta, \varphi | z)) = \mathbb{E}_{\hat{Q}(z)}[-\log p(\mathcal{X}, z, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{Q}(z)) \tag{44}$$

As a result, the bound in GD-based CVB of the CVB-LGDA can be expressed as :

$$-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq \hat{\mathcal{F}}(\hat{Q}(z)) = \mathbb{E}_{\hat{Q}(z)}[-\log p(\mathcal{X},z,c|\varepsilon,\zeta)] - \mathcal{H}(\hat{Q}(z)) \tag{45}$$

$$\hat{\mathcal{F}}(\hat{Q}(z)) \leq \tilde{\mathcal{F}}(\tilde{Q}(z)) \triangleq \min_{\tilde{Q}(\theta)\tilde{Q}(\varphi)} \tilde{\mathcal{F}}(\tilde{Q}(z)\tilde{Q}(\theta)\tilde{Q}(\varphi)) \tag{46}$$

The Eq. 105 shows the GD-based CVB being a better and improved approximation than the standard VB after the parameters are marginalized out in the collapsed space of the latent variables. In addition, minimizing the variational free energy $\hat{\mathcal{F}}(\hat{Q}(z))$ in Eq. 104 with respect to ψ_{ijk} leads to the multinomial update in each class as shown in Eq. 107. Using [89], we have:

$$\log Q_j(z_j) \propto \mathbb{E}_{i\neq j}[\log p(\mathcal{X},z)] \tag{47}$$

where $\mathbb{E}_{i\neq j}[...]$ is the expectation with respect to the variational distribution Q for all z_i such that $i \neq j$. The equation above leads to:

$$Q_j(z_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\mathcal{X},z)])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\mathcal{X},z)])dz_j} \tag{48}$$

For our classification problem (in the collapsed space) with z being discrete, we get:

$$\hat{\psi}_{ijkc} = \hat{Q}(z_{ij}=k|c) = \frac{\exp(\mathbb{E}_{\hat{Q}(z^{-ij})}[\log p(\mathcal{X},z^{-ij},z_{ij}=k,c|\varepsilon,\zeta])}{\sum_{k'=1}^{K}\exp(\mathbb{E}_{\hat{Q}(z^{-ij})}[\log p(\mathcal{X},z^{-ij},z_{ij}=k',c|\varepsilon,\zeta])} \tag{49}$$

In the GD-based CVB, the latent variables are sampled from the variational posterior distribution $\hat{Q}(z)$ and uses the GD based-CGS. The expected topic assignments lead to parameters estimations when the Markov chain is stationary. These conclusions are also found in [30].

2.3.3.6 Gaussian Approximation in GD-CVB: Second order Taylor approximation

For large datasets, the implementation of the GD- based CVB in the CVB-LGDA, even though accurate is very expensive as it computes several expectations similar to Dirichlet-based CVB in [30]. Dealing with this problem requires the use of Gaussian approximations to estimate the multinomial parameter $\hat{\psi}_{i'jkc}$ and speed up the process. In this scheme of improving the speed, the counts in the Gibbs sampler act as fields and can be defined as a large sum of independent Bernoulli variables $\vec{1}(z_{i'j}=k)$, each with parameter $\hat{\psi}_{i'jkc}$ as shown in [30]. So, the mean of the sum of the Bernoulli variables means and variance of the sum of the Bernoulli variable variances [30] are respectively computed as :

$$\mathbb{E}_{\hat{Q}}[N_{jkc.}^{-ij}] = \sum_{i'\neq i}\hat{\psi}_{i'jkc} \tag{50}$$

$$Var_{\hat{q}}[N_{jkc.}^{-ij}] = \sum_{i'\neq i}\hat{\psi}_{i'jkc}(\vec{1}-\hat{\psi}_{i'jkc}) \tag{51}$$

The variance and the mean are then used in the Gaussian approximation to estimate the expected values of logarithmic expressions such as $E_{\hat{Q}}[\log(\alpha + N_{jk.})]$. Using [30], we obtained:

$$\mathbb{E}_{\hat{Q}}[\log(\alpha + N_{jkc.})] \approx \log(\alpha + \mathbb{E}_{\hat{Q}}[N_{jkc.}]) - \frac{Var_{\hat{Q}}[N_{jkc.}]}{2(\alpha + \mathbb{E}_{\hat{Q}}[N_{jkc.}])^2} \tag{52}$$

Therefore, the expression above becomes:

$$\exp(\mathbb{E}_{\hat{Q}}[\log(\alpha + N_{jkc.})]) \approx (\alpha + \mathbb{E}_{\hat{Q}}[N_{jkc.}]) - \exp\left(\frac{Var_{\hat{Q}}[N_{jkc.}]}{2(\alpha + \mathbb{E}_{\hat{Q}}[N_{jkc.}])^2}\right) \tag{53}$$

This is the second-order Taylor expansion used as an approximation [4]. The model computes an extremely large amount of expectations; so the scheme is found to be very useful in speeding up the GD-CVB algorithm. The GD-based CVB in CVB-LGDA update is finally expressed as :

$$\begin{aligned}
&\hat{Q}(z_{ij} = k|c) = \hat{\psi}_{ijkc} \propto \\
&\left[\frac{(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}^{-ij}])}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}^{-ij}])}\right] \\
&\times \left[\frac{(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}])(\eta_\nu + \sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])}{(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])}\right] \\
&\times \exp\left(-\frac{Var_{\hat{Q}}(N_{jk.}^{-ij})}{2(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
&\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{2(\beta_{ck} + (\sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}^{-ij}])^2}\right) \\
&\times \exp\left(-\frac{Var_{\hat{Q}}((N_{.k\nu_{ij}}^{-ij})}{2(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}])^2}\right) \\
&\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{2(\alpha_{ck} + \beta_{ck} + \mathbb{E}_{\hat{Q}}[\sum_{l=k+1}^{K+1} N_{jl.}^{-ij}])^2}\right) \\
&\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{d=\nu+1}^{V+1} N_{.kd_{ij}}^{-ij})}{2(\eta_\nu + (\sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])^2}\right) \\
&\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{d=\nu}^{V+1} N_{.kd_{ij}}^{-ij})}{2(\lambda_\nu + \eta_\nu + (\sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[n_{.kd_{ij}}^{-ij}])^2}\right)
\end{aligned} \tag{54}$$

This equation shows that CVB-LGDA samples its latent variables from a variational posterior distribution Q in the collapsed space of latent variables.

2.3.3.7 Parameters estimates: Predictive distributions

The CVB-LGDA's generative process for an unseen document (image,3D object, or a video frame) requires its predictive distribution expressed in terms of its parameter θ_j conditional on the model hyperparameters $(\varepsilon, c) = (\alpha_c, \beta_c)$. Using [30], document parameter distribution is given as:

$$\hat{\theta}_{jk} = \frac{(\alpha_{kc} + \mathbb{E}_Q[N_{jk.}])(\beta_{kc} + \sum_{l=k}^{K+1} \mathbb{E}_Q[N_{jk.}])}{(\alpha_{kc} + \beta_{kc} + \sum_{l=k}^{K+1} \mathbb{E}_Q[N_{jk.}])} \tag{55}$$

Conditional on the topic k, the predictive distribution of the words is expressed as φ_{kw} such that:

$$\hat{\varphi}_{kw} = \frac{(\lambda_v + \mathbb{E}_Q[N_{.kv_{ij}}])(\eta_v + \sum_{d=v+1}^{V+1} \mathbb{E}_Q[N_{.kd_{ij}}])}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} \mathbb{E}_Q[N_{.kd_{ij}}])} \tag{56}$$

2.3.4 Empirical likelihood: Evaluation method for the topic model

Very often, the lack of reliable topic labels for the dictionary codewords leads to the need for an evaluation method to assess or validate the robustness of the estimated topic model [18]. The goal is to compute efficiently the probability of the held-out dataset [42, 18]. After estimation of the predictive distributions, we used the empirical likelihood estimate scheme presented in [18] as a validation method. In the CVB-LGDA model, the likelihood [30, 18] could be reduced to:

$$p(\mathcal{X}_{unseenDoc}) = p(\mathcal{X}_{unseenDoc}|c, \varepsilon, \zeta) = \prod_{ij} \sum_{k} \hat{\theta}_{jk} \hat{\varphi}_{kw} \tag{57}$$

such that the counts $\mathbb{E}_Q[N_{jk.}]$, $\mathbb{E}_Q[N_{.kv_{ij}}]$, and $\mathbb{E}_Q[N_{.kd_{ij}}]$ of the unseen document are obtained from the GD-based CVB sampling process in the collapsed space. The parameters of the unseen document (or its codewords and topic distributions) are then used to predict its likelihood.

The classification problem is also reduced to a likelihood estimation approach which approximates the distribution of codewords in each class. It evaluates the topic model in each class [18]. It is designed to predict the likelihood of the unknown document. Therefore, the predictive likelihood $p(\mathcal{X}|c, \varepsilon, \zeta)$ is estimated as follows: for an unseen document to be classified, some pseudo documents are generated with parameters θ using the GD priors from the training set. Once we obtain the best candidates of documents in each class, we estimate their word probability distribution given the corpus parameter φ which leads to the class conditional probability $p(\mathcal{X}|c, \varepsilon, \zeta)$. With the class conditional probability, we can assess the probability of seeing the test set (unknown document) in the class. The class label is then given to the unseen document if it has the highest likelihood. The scheme is similar to [18, 2]. The empirical likelihood estimate is assumed to be robust compared to a topic model's perplexity scheme as an evaluation method (validation) of the performance of the model.

2.3.5 Bayesian decision boundary for classification

The empirical likelihood estimate provides the probability of seeing the test set. In this classification problem Fig. 2.2, it is used to assess the class of the test set where the probability of seeing its class is proportional to the class likelihood for a uniform class prior. Consequently, once the model parameters and latent variables are estimated for the generative process in each class, then given an unseen document (image, 3D object, face expression, video frame) with its BoW representation $\mathcal{X}$, the probability of each class label (predictive model) is expressed as:

$$p(c|\mathcal{X}, \mu, \varepsilon, \zeta) \propto p(\mathcal{X}|c, \varepsilon, \zeta)p(c|\mu) \propto p(\mathcal{X}|c, \varepsilon, \zeta) \tag{58}$$

As a result, to assign a category to an unseen document, the decision is ultimately made by the category label with the highest likelihood probability [2] such that:

$$C^* = \underset{c}{\operatorname{argmax}}\, p(\mathcal{X}|c, \varepsilon, \zeta) \tag{59}$$

2.3.6 Model Selection

It is really challenging in topic modeling framework to choose and fix the number of topics. As already explained in [5], two reasons tend to justify this tremendous handicap: the difficulty in selecting an appropriate criterion is one reason as it has been said that an optimization scheme with respect to the criterion could be very expensive in topic modeling. The second reason is that data or document collections do grow over time, and the database tends to contain entities (topics, codewords) and structures that are new or different from the original training set elements. As a result, this is a serious drawback in the process of providing a better generalization of the model to future or unseen data. As we are working in the finite dimensional space using finite mixtures where we deal with finite number of topics, and fixed size in the vocabulary, our option for a model selection has been to implement an exhaustive search which ultimately takes into account a series of number of topics along with vocabulary sizes in search for the optimal values (number of topics, and vocabulary size) that provide the highest classification accuracy rate. In other words, this scheme despite being expensive is an attempt to provide the optimal number of topics and vocabulary size for a better description of our topic model.

2.3.7 Comparison with other classification models

To evaluate our proposed model and inference technique, we set up a goal to compare the new approach with the following traditional models used in classification framework.

2.3.7.1 K-Nearest Neighbor

As one of the simplest algorithm in machine, the The K-nearest Neighbor (KNN) has been widely used in machine learning especially in classification. In categorization (supervised learning), the unseen object is classified by a majority vote of its nearest neighbors. As a result, the unseen object is assigned to the class label most common within its k-nearest neighbors. When k=1 the object is just assigned to the class of its single nearest neighbor. Very often the KNN uses a variety of distance metrics such as the Euclidean metric Correlation distance, Mahalanobis distance, Mahalanobis distance, Mahanattan distance, etc [43, 44].

2.3.7.2 Support Vector Machine

In categorization approaches, the Support Vector Machine (discriminative model) is widely used. While it can perform linear classification, the scheme has ability to perform also in a non-linear classification framework using the kernel trick that maps the inputs into high dimensional spaces. The core idea behind the SVM is to construct a hyperplane (set of hyperplanes) in these high dimensional space where a good separation is reached when the hyperplane that has the largest functional margin; in other words the distance to the nearest training points of any class in (Correlation Kernels sor Support vector Machines Classification with applications in cancer Data) Traditional Kernels include RBF (radial basis functions), linear, hyperbolic tangent, and polynomial. As pointed out in [44], one of the challenges in using SVM is the choice of the Kernel as it dictates the performance of the classifier with the dataset. This has resulted in the use of appropriate kernels within the bag of word framework such as the Hellinger Kernel, Histogram intersection kernel, Generalized Gaussian Kernel (Eq. 62) where D can be Euclidean or Chi-square distance;

A is a scaling parameter that can be determined through cross-validation [45, 46]. These Kernels are defined as:

- Hellinger Kernel:

$$K(h_1, h_2) = \sum_{i=1}^{N} \sqrt{h_1(i) h_2(i)} \tag{60}$$

- Histogram intersection Kernel:

$$I(h_1, h_2) = \sum_{i=1}^{N} \min(h_1(i), h_2(i)) \tag{61}$$

- Generalized Gaussian Kernel:

$$K(h_1, h_2) = \exp\left(\frac{-D(h_1, h_2)^2}{A}\right) \tag{62}$$

  D can be Euclidean distance (leading to the RBF Kernel), χ^2 distance etc.

$$D_{\chi^2}(h_1, h_2) = \sum_{i=1}^{N} (\frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}) \tag{63}$$

- Quadratic distance (cross bin):

$$K(h_1, h_2) = \sum_{i,j}^{N} A_{ij} \left(h_1(i) - h_2(j)\right)^2 \tag{64}$$

2.3.7.3 Backpropagation Neural Network

Used in artificial Neural network (ANN) for classification, the backpropagation neural network (BPNN) ultimately approximates the nonlinear relationship between the input and its ouputs through adjustment in the weigth values. Its goal is to minimize the error function, using gradient descent scheme so that output value matches the target value. Though, in this scheme, the error calculated at the output is propagated back to its layers: a feedforward and a backpropagation processes are executed one after the other until the error between the target value and the network output (at the output layer) is fully minimized. One of the challenges in BPNN is about determining the number of neurons in the hidden layers [43].

2.3.7.4 Generative adversarial Networks

The research and science community have seen a recent interest in the generative models especially in the Generative Adversarial Networks (GANs) [47] due to its wide applications including classification using deep learning. Basically, the model is a combination of a generative model and a discriminative one that compete against each other in a zero-sum game framework. In fact, the generator takes noise from input and generates samples from it while the discriminator tries to distinguish samples from the generator and the original training set. In this game, the generator finally learns to generate very realistic samples whereas the discriminator becomes also very talented in distinguishing the generated data from the real data. The ultimate goal is to make the generator generate samples that become indistinguishable from the real data.

	Images	Face expressions	3D	Action Recognition in videos
LGDA	55.28%	70.4%	61%	68%
GD-LDA	65.1%	69%	56.23%	51%
LDA	57%	50.3%	54%	50.25%
CVB-LDA	59.6%	61.40%	60.57%	60.46%
CVB-LGDA	70.27%	89.8%	63.46%	70.12%

Table 2.3: Comparison between the new CVB-LGDA model and the other schemes within the BoW framework

2.4 Experimental results

In the topic modeling literature, several applications have often focused on text modeling. In our experiments in this chapter, we are implementing some challenging applications to show the merits of the new approach. These applications ultimately include: image and 3D object classification, facial expressions recognition and their categorization, and action recognition in videos. Following the bag of visual words framework, these applications in this chapter mainly emphasize on representations using local features.

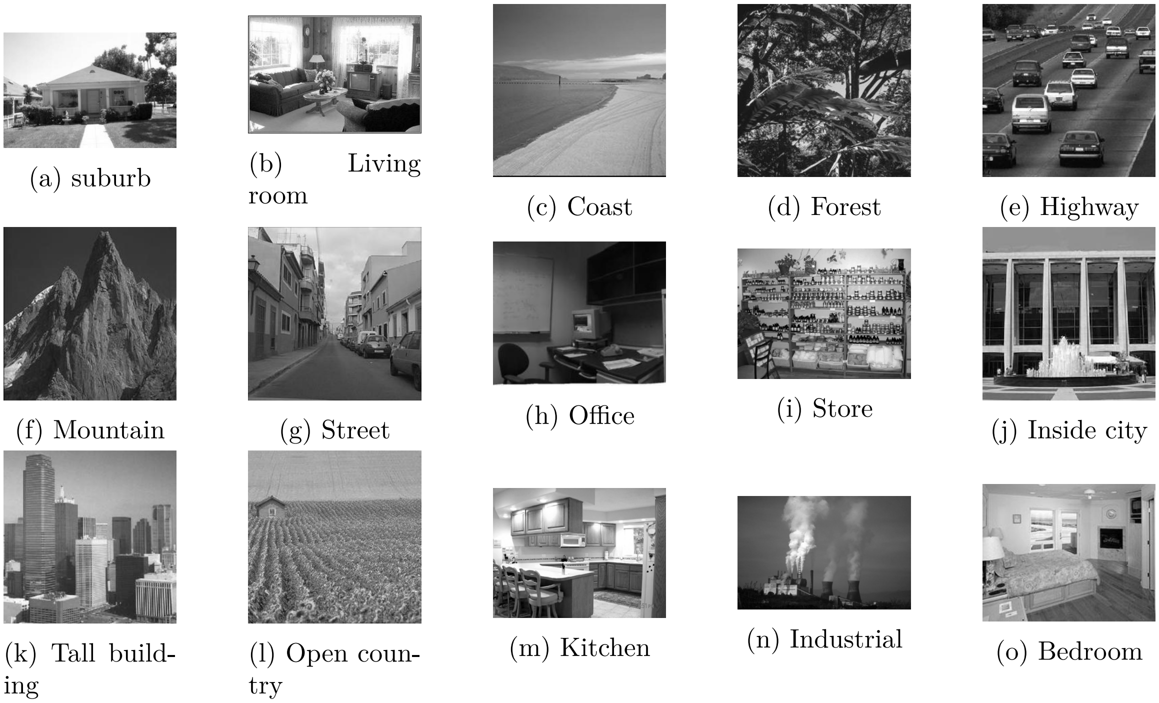
2.4.1 Image Categorization

2.4.1.1 Methodology

In our experiments, we constructed our model using the well-known grayscale 15 categories natural scenes dataset [48] . As illustrated in Fig. 2.3 and Table 3.3, this widely known and challenging data set includes the folloiwng categories suburb, living room, coast, forest, highway, mountain, street, office, store, bedroom, inside city, tall building, open country, kitchen, and industrial. In each category, the data is subdivided into two parts: the testing set contains 100 samples while the remaining constitutes the training set.

In the BoW framework, the local feature representation of the corpus leads to vectors of counts in each document (image) in the preprocessing stage. The following steps are essential in the BoW representation: first, using the entire collection of the corpus, local features (from local patches) are extracted from them using the SIFT (Scale Invariant Feature Transform) algorithm (Fig. 3.12). The collection of the training set image descriptors is clustered using K-means algorithm to find a unique representation in the dataset (where similar patches are grouped together to form a cluster). After quantization, each cluster center is a codeword and the total number of codewords is the codebook (dictionary or vocabulary). With the codebook, each image (document) is then represented as a vector of counts: this is the bag of visual word representation of the corpus.

The training set count data are then used to implement the CVB-LGDA model with asymmetric GD priors. The topic parameters estimation leads to the predictive model. Using the topic predictive distributions, we used the empirical likelihood framework as evaluation method for the robustness of the topic distribution. It then leads to the estimation of the class likelihood (class conditional probability). The class conditionals help predicting the class label of unseen images or documents. As a result, the category of



(a) suburb

(b) Living room

(c) Coast

(d) Forest

(e) Highway

(f) Mountain

(g) Street

(h) Office

(i) Store

(j) Inside city

(k) Tall building

(l) Open country

(m) Kitchen

(n) Industrial

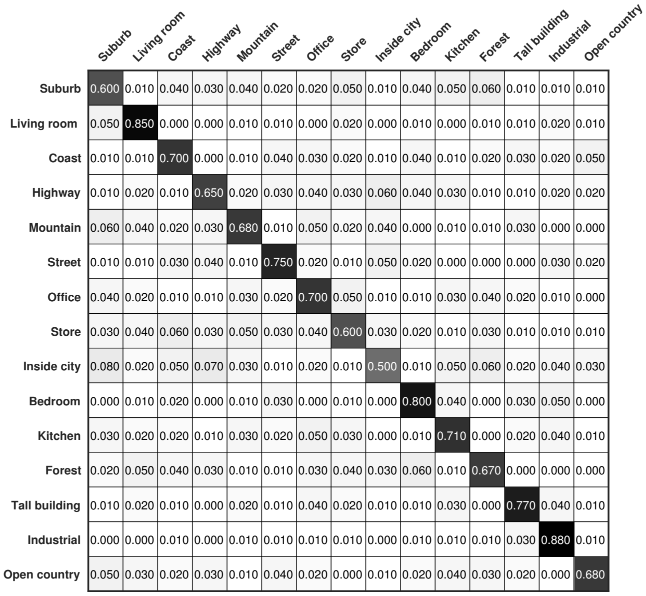
(o) Bedroom

Figure 2.3: Examples from the natural scenes images dataset (15 categories).

Categories	Size
suburb	241
living room	289
cost	360
forest	328
highway	260
mountain	374
street	292
office	215
store	315
Bedroom	216
Inside City	308
Tall buidling	356
Open country	410
Kitchen	210
Industrial	311

Table 2.4: size of each image category.

unseen image is chosen by the class with the highest class posterior distribution which is equivalent to the class conditional probability for a uniform prior.



	Suburb	Living room	Coast	Highway	Mountain	Street	Office	Store	Inside city	Bedroom	Kitchen	Forest	Tall building	Industrial	Open country
Suburb	0.600	0.010	0.040	0.030	0.040	0.020	0.020	0.050	0.010	0.040	0.050	0.060	0.010	0.010	0.010
Living room	0.050	0.850	0.000	0.000	0.010	0.010	0.000	0.020	0.000	0.010	0.000	0.010	0.010	0.020	0.010
Coast	0.010	0.010	0.700	0.000	0.010	0.040	0.030	0.020	0.010	0.040	0.010	0.020	0.030	0.020	0.050
Highway	0.010	0.020	0.010	0.650	0.020	0.030	0.040	0.030	0.060	0.040	0.030	0.010	0.010	0.020	0.020
Mountain	0.060	0.040	0.020	0.030	0.680	0.010	0.050	0.020	0.040	0.000	0.010	0.010	0.030	0.000	0.000
Street	0.010	0.010	0.030	0.040	0.010	0.750	0.020	0.010	0.050	0.020	0.000	0.000	0.000	0.030	0.020
Office	0.040	0.020	0.010	0.010	0.030	0.020	0.700	0.050	0.010	0.010	0.030	0.040	0.020	0.010	0.000
Store	0.030	0.040	0.060	0.030	0.050	0.030	0.040	0.600	0.030	0.020	0.010	0.030	0.010	0.010	0.010
Inside city	0.080	0.020	0.050	0.070	0.030	0.010	0.020	0.010	0.500	0.010	0.050	0.060	0.020	0.040	0.030
Bedroom	0.000	0.010	0.020	0.000	0.010	0.030	0.000	0.010	0.000	0.800	0.040	0.000	0.030	0.050	0.000
Kitchen	0.030	0.020	0.020	0.010	0.030	0.020	0.050	0.030	0.000	0.010	0.710	0.000	0.020	0.040	0.010
Forest	0.020	0.050	0.040	0.030	0.010	0.010	0.030	0.040	0.030	0.060	0.010	0.670	0.000	0.000	0.000
Tall building	0.010	0.020	0.010	0.000	0.020	0.010	0.040	0.020	0.010	0.010	0.030	0.000	0.770	0.040	0.010
Industrial	0.000	0.000	0.010	0.000	0.010	0.010	0.010	0.010	0.000	0.010	0.010	0.010	0.030	0.880	0.010
Open country	0.050	0.030	0.020	0.030	0.010	0.040	0.020	0.000	0.010	0.020	0.040	0.030	0.020	0.000	0.680

Figure 2.4: Confusion matrix for the natural scenes classification problem.

2.4.1.2 Results

The CVB-LGDA was able to provide a better result in terms of accuracy as shown in the confusion matrix (Fig. 3.2). In model selection (Fig. 2.6), the optimal number of topics obtained is $K = 145$ while the optimal vocabulary size is $V = 1450$. The overall accuracy rate is 70.27% at these optimal values. Due to an efficient feature representation, these results ultimately show the flexibility of the new approach (robust prior) as the model has ability to compute true posterior distributions rather than approximating them as in variational methods with the variational posterior distributions. In addition, a correlation map (Fig. 3.3) shows the dependency between any two classes in our categorization problem. These results reinforce the concept of generalization of the LDA model (to different data types) in which richer codewords, robust generative schemes (with flexible priors), and inference techniques could enhance performance.

2.4.2 Facial Expression recognition

Facial expressions and emotions recognition are getting a lot of attention today as they are hot topics in data analytics due to the impact of social media (Twitter, Instagram, Facebook, Flickr, and Youtube). The facial expression model is concerned with a visual learning process that can also focus on the classification of characteristics such as facial motions used in various applications (image understanding, virtual reality, synthetic face animation, facial nerve grading in medicine etc [49, 50]).

In this application, we decided to use a very flexible and robust descriptor from the Fast LBP-TOP (Local Binary Patterns histogram from Three Orthogonal Planes) scheme as suggested in [51] for facial expression images modeling. We considered the JAFFE (Japanese Female Facial Expression) dataset (See Figs. 3.4 and 3.5). It contains 213 images obtained from 10 Japanese females showing 7 facial expressions such as surprise, anger, happiness,

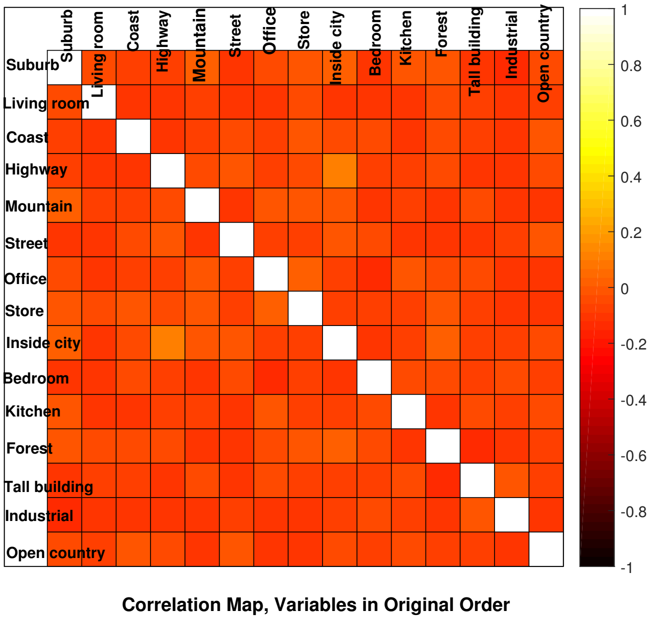


Figure 2.5: Natural scene images correlation map.

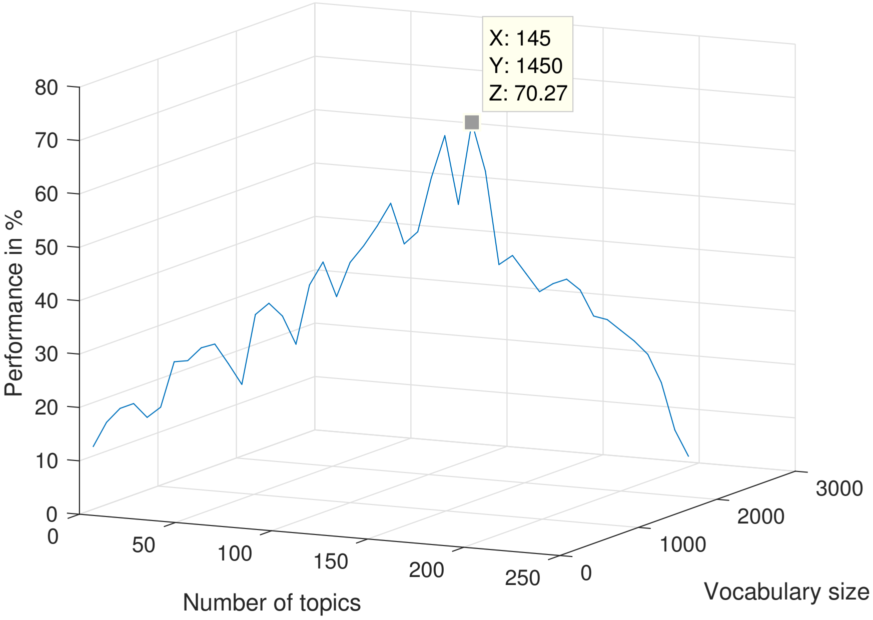


Figure 2.6: Optimal number of topics and vocabulary size for image classification problem.

sadness, fear, disgust, and neutral. The first task is to group these females according to these seven expressions representing our different classes. The dataset is partitioned into a training set and a testing set. From the training set, we obtained the corpus features from the Fast LBP-TOP descriptors. These normalized histograms are then clustered and then quantized to get the codebook of the corpus leading to the bag of visual word representation of images (documents) in the training set. Prior to the BoW representation of the corpus, key features are drawn from each image regions of interest (Fig. 3.6). Within the BoW, the documents with vectors of counts are then used to build the CVB-LGDA model where we compute the parameters of the topics in each class; and then use the topic distribution in each class to predict the category of unseen documents. As a result, the class label is given to the class with the highest posterior distribution or class conditional probability (for a uniform class prior).

The confusion matrix (Fig. 3.7) obtained shows high accuracy rate of 89.8% as shown in Fig. 2.12. which outperforms its competitors (see Table 3.2). In addition, the optimal number of topics is $K = 70$ while the optimal vocabulary size is $V = 105$. We illustrated a correlation map (Fig. 2.11) that measures the dependency between any two categories in this classification problem. It also demonstrates the capability of the GD in coping with both negatively and positively correlated data.

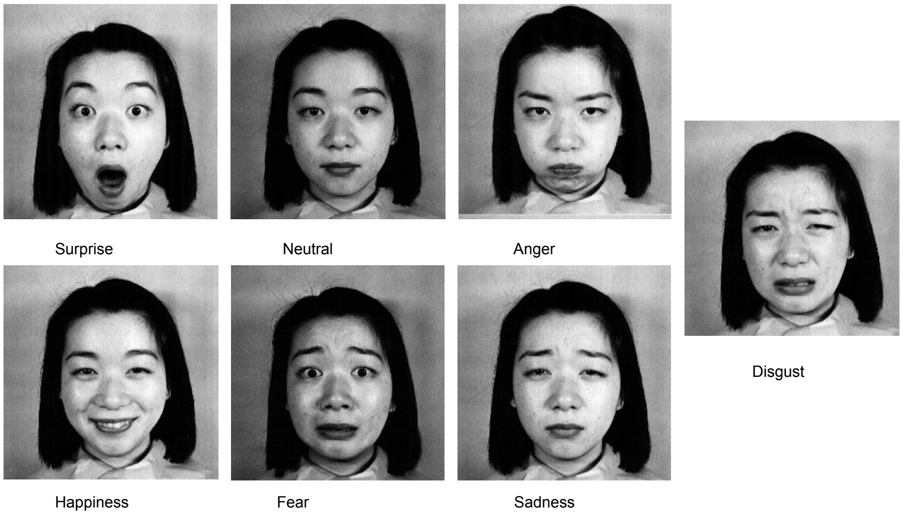


Figure 2.7: Facial expressions and emotions in the JAFFE dataset

2.4.3 3D object classification

The dataset (Fig. 2.18) we consider in this application contains 10 classes of 3D objects [52]. These classes are: stapler, car bicycle, head, computer, mouse, toaster, cellphone, shoe, and iron. It is important to point out that these are collections of objects under different 2D views to implicitly create a 3D concept of the objects (the bicycle for instance) as illustrated in Fig. 2.18. For the training set, 7 (3D) objects are randomly selected with around 250 images per 3D object. The remaining is allocated to the testing set in each class. We obtained around 80 images per object.

From observation in the dataset, in every 3D class, the characteristics of the object are represented using a very large collection of the object's 2D images seen from different

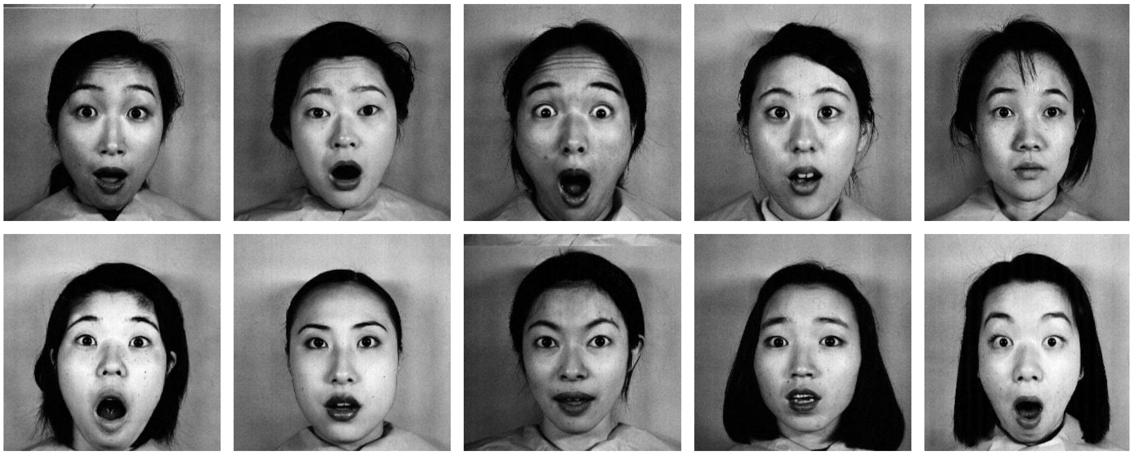


Figure 2.8: Women showing a "surprised" facial expression

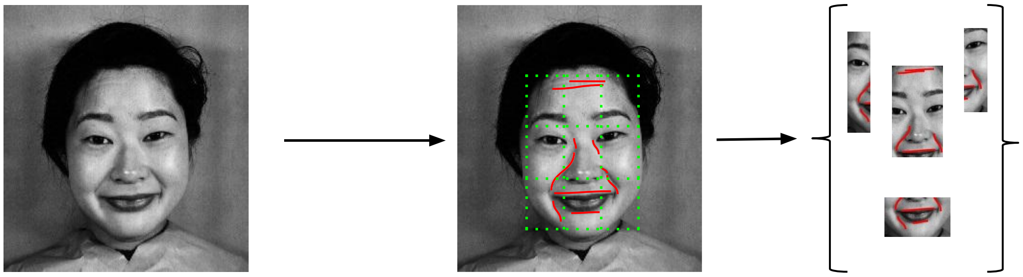
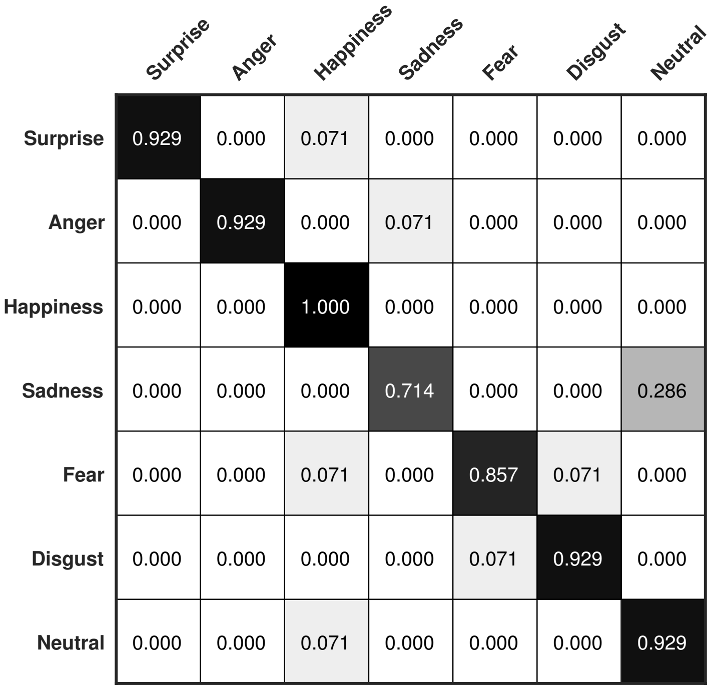


Figure 2.9: Facial Expression: Key Regions of Interest and Extraction

angles or views. In other words, these views are used to generate the 3D characteristics of the object in each class. As a result, constructing a 3D class is equivalent to extracting the features characteristic from its different parts emphasized by the different 2D views. In this application, this is also done using the 2D SIFT descriptors so that each 3D object class contains its intrinsic bag of features (Fig. 2.16). The entire collection of features from the 3D object classes is first clustered using K-means and then quantized to obtain the codebook of the corpus. The codebook provides the BoW representation (count data) of each 3D class. The data is then used to implement the CVB-LGDA which preforms a classification's task based on the topic signatures from every 3D class. With the flexibility of the GD prior, the model could easily cope with a large vocabulary size and an increasing number of topics in the dataset.

The optimal number of topics obtained for 3D object modeling is $K = 180$ for an optimal vocabulary size of $V = 1800$. At these optimal values (Fig. 2.15), the accuracy rate shown by the confusion matrix (Fig. 2.13) reaches a maximum of 63.46%. Due to the high level of noise (background) in the 2D images representing the 3D objects as shown in the example in Fig. 2.18, we can say this is a very satisfactory result also taking into account the complexity in the overall 3D dataset structure in comparison to the image categories data. The robustness can be compared to the other models as illustrated in Table 3.2. The model was still able to provide a better result with a very challenging dataset where correlation analysis has been useful as shown in Fig. 2.14.



	Surprise	Anger	Happiness	Sadness	Fear	Disgust	Neutral
Surprise	0.929	0.000	0.071	0.000	0.000	0.000	0.000
Anger	0.000	0.929	0.000	0.071	0.000	0.000	0.000
Happiness	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Sadness	0.000	0.000	0.000	0.714	0.000	0.000	0.286
Fear	0.000	0.000	0.071	0.000	0.857	0.071	0.000
Disgust	0.000	0.000	0.000	0.000	0.071	0.929	0.000
Neutral	0.000	0.000	0.071	0.000	0.000	0.000	0.929

Figure 2.10: Confusion matrix from the Facial expressions classification

2.4.4 Action recognition in videos

A robust motion recognition system and a deep analysis represent the two best ingredients for a complete implementation of human behaviour's understanding using automated surveillance systems [53]. In this chapter, the action recognition of motions in video has been implemented with the optical flow algorithm which helps collecting relevant features for the BoW representation of the corpus data in order to build our model. In this experiment, we have used the KTH dataset which contains 2391 video sequences at 25 frames [54, 55]. It mainly includes individuals (25 actors) in 4 scenarios performing 6 types of human actions (walking, running, jogging, boxing, hand waving, and hand clapping) as illustrated in Table 3.6. In these figures, each column represents a human action in 4 different scenarios. For processing purpose, the sequences were downsampled to a resolution of 160 by 120 pixels with a length of 4 seconds.

In our experiment, 60% of the dataset were used for training while the remaining constitutes the testing set. Around 100 frames were collected from each video sequence in each class.

Within the BoW, we first needed a method that could capture the motion of objects in the video sequences for a better representation of the dataset. And this is obtained with the optical flow scheme proposed by the Horn and Schunck algorithm [56]. It is a global approach that has ability to yield a dense flow often needed and preferred in computer vision applications.

After obtaining the optical flow for the frames (images), a threshold is set to only recover the most relevant components of the optical flow matrices. These relevant components of all

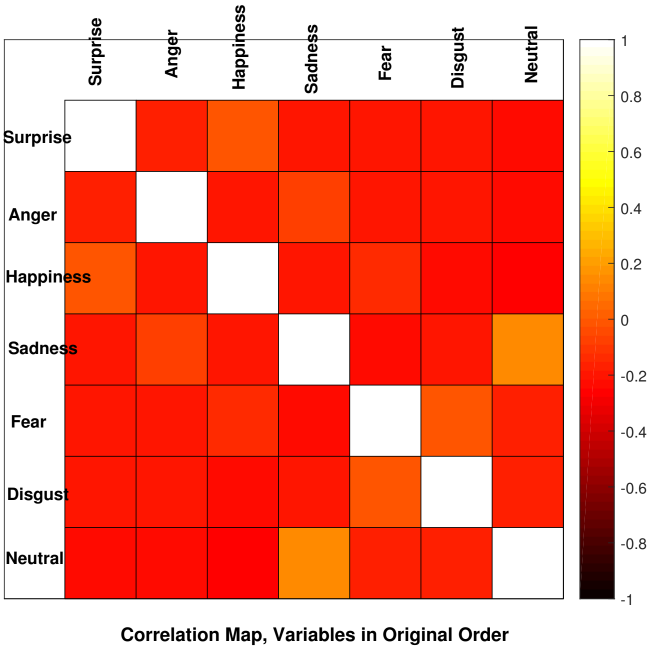


Figure 2.11: Correlation map for facial expression categories

categories of actions in the training set are then grouped and then clustered with a K-means algorithm in order to express a unique representation as a codebook. From the codebook, each component can be represented as a BoW feature similar to [35], which is used in our CVB-LGDA model.

This model with the optical flow technique is very computationally expensive as it requires so many features; however, it was able to provide an overall accuracy of 70.12%. The stability of the model insured the motion detection, recognition and classification in the video sequences. This is also due to the efficiency in the GD prior within the collapsed variational Bayes inference scheme.

From the results obtained in these applications (Table 3.2), we can say that the CVB-LGDA model is very robust and could be definitely an alternative to finite mixture models considering its performances [57, 7].

It is important to finally observe that as the global method proposed by Horn and Schunk has some limitations due to the very sensitiveness of the optical flow algorithm to noise, an improvement could be a framework that combines the local methods (robust to noise) proposed by Kanade and Lucas and the global schemes of Horn-Schunck's approach (dense flow fields). This hybrid scheme should ultimately provide the best optical flow features.

2.4.5 Classification results with other supervised models

In our setting, a 5-fold cross-validation scheme has been implemented in the classification models. And to ensure stability in the results the cross-validation technique has been performed 8 times where finally the classification accuracy was then measured as the averaged accuracy over these 8 runs.

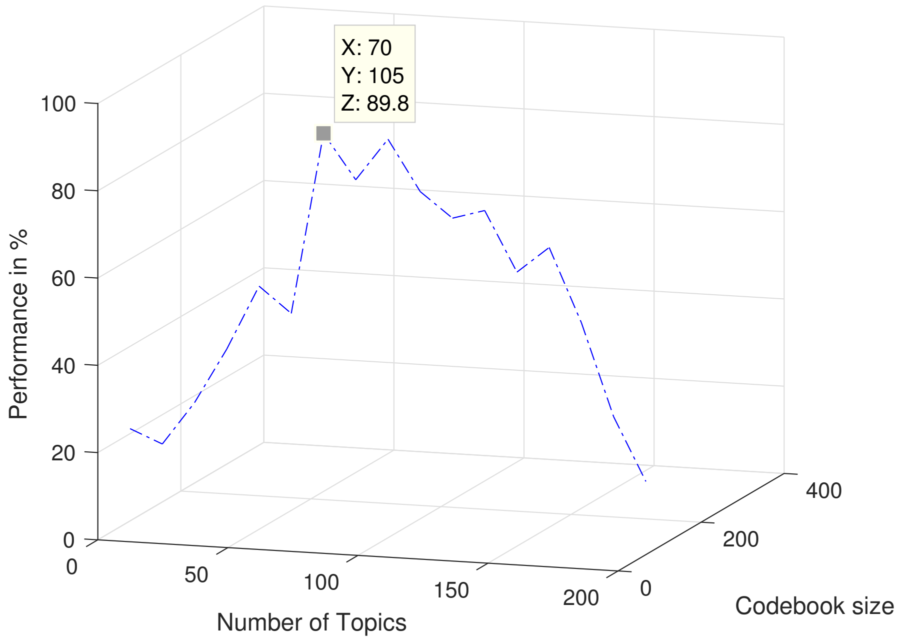


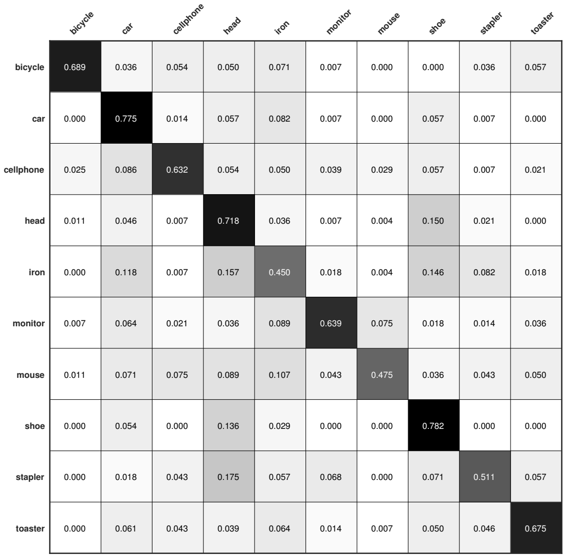
Figure 2.12: Model selection for facial expressions

Each table summarizes the three classification models (KNN, SVM, and BPNN) to each dataset. In this chapter, as our entire collections have a BoW feature representation, the distance of choice in case of the KNN was the Euclidean distance. We therefore performed the K-nearest neighbor algorithm on different datasets such as images, videos, 3Ds. We used different values of K to analyze the influence it has on the performance of the classifier. As a result, values such as $K = 1$, $K = 7$, and $K = 10$ have been selected. The different average accuracy values obtained from these datasets are summarized in Tables 2.5, 2.6, 2.7, and 2.8. From this table, we can observe (through the performance of the model using these datasets) that the best results were obtained at lower values of K ($K = 1$ and $K = 7$). Any value of K above 7 has seen a decrease in the performance. In addition, For all datasets in low dimension, KNN provides good performance than in the case of high dimensional data (videos and 3Ds) due the large vocabulary size. Among these 3 classifiers the high performances were obtained with face expression dataset (69.4%), natural scene (63.14%), and videos (66.1%).

In SVM, we decided for the Radial Basis Function (RBF) Kernel. For this Kernel, the parameter A is taken from $\{0.1, 1.0, 4\}$. The results in terms of averaged classification accuracy obtained in Tables 2.5, 2.6, 2.7, and 2.8 show that the performance hits a ceiling at $A = 1$ and from that point, we notice a significance decrease in the performance.

From these datasets, the videos (activity recognition in videos) and the images datasets (scenes and face expressions) provided the best results: 68.1%, 66.4%, and 71.25%, respectively. Though, their performances has dropped when the value of A increased.

In the case of BPNN, we first equipped the hidden layer with 4 neurons and then 6 neurons. The output layer carries neurons equal to the total number of categories in our classification problem. We observed that in our neural network model, the accuracy increases with the number of neurons in the output layer. Though, everything is getting



	bicycle	car	cellphone	head	iron	monitor	mouse	shoe	stapler	toaster
bicycle	0.689	0.036	0.054	0.050	0.071	0.007	0.000	0.000	0.036	0.057
car	0.000	0.775	0.014	0.057	0.082	0.007	0.000	0.057	0.007	0.000
cellphone	0.025	0.086	0.632	0.054	0.050	0.039	0.029	0.057	0.007	0.021
head	0.011	0.046	0.007	0.718	0.036	0.007	0.004	0.150	0.021	0.000
iron	0.000	0.118	0.007	0.157	0.450	0.018	0.004	0.146	0.082	0.018
monitor	0.007	0.064	0.021	0.036	0.089	0.639	0.075	0.018	0.014	0.036
mouse	0.011	0.071	0.075	0.089	0.107	0.043	0.475	0.036	0.043	0.050
shoe	0.000	0.054	0.000	0.136	0.029	0.000	0.000	0.782	0.000	0.000
stapler	0.000	0.018	0.043	0.175	0.057	0.068	0.000	0.071	0.511	0.057
toaster	0.000	0.061	0.043	0.039	0.064	0.014	0.007	0.050	0.046	0.675

Figure 2.13: 3D object confusion matrix

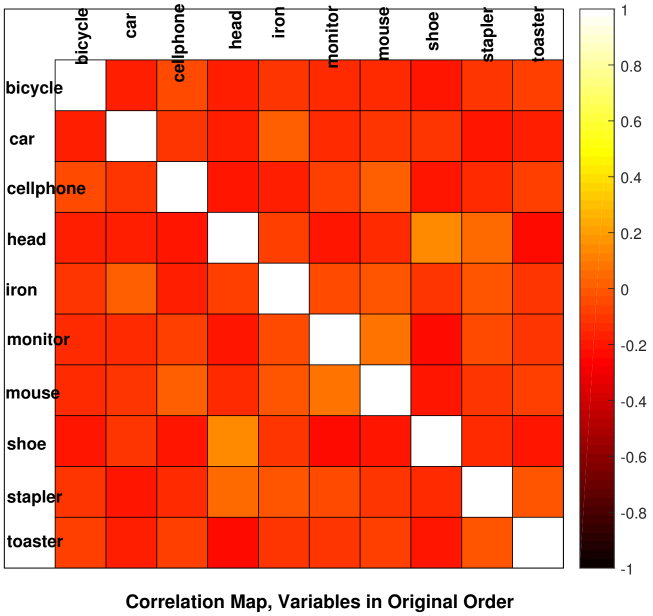


Figure 2.14: Correlation map for the 3D objects categories.

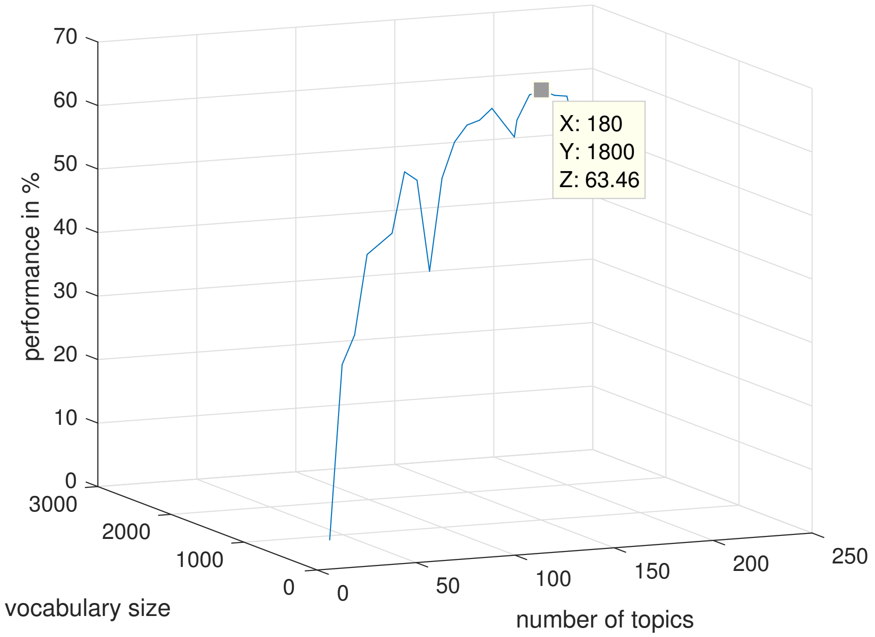


Figure 2.15: Optimal number of topics and vocabulary size for 3D modeling

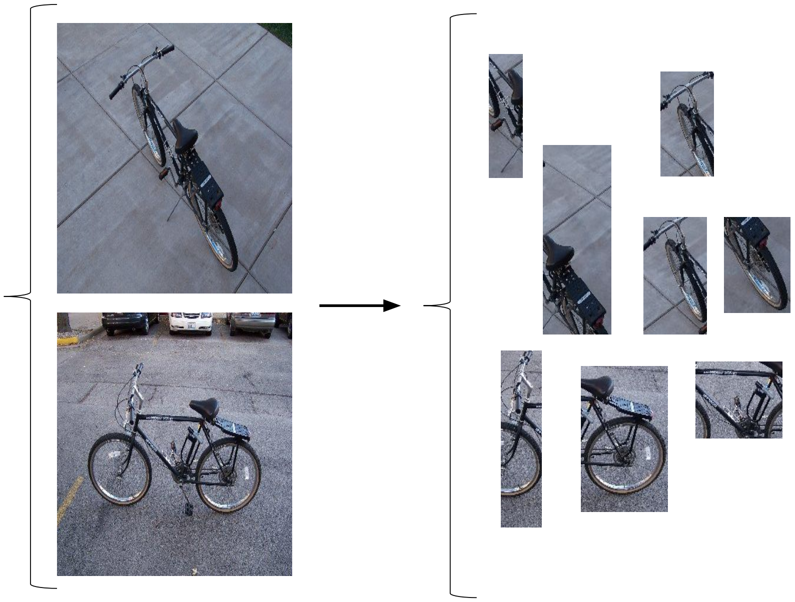


Figure 2.16: 2D Features extraction for a 3D modeling

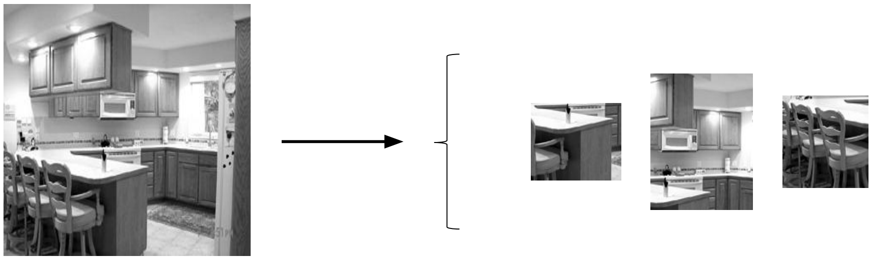
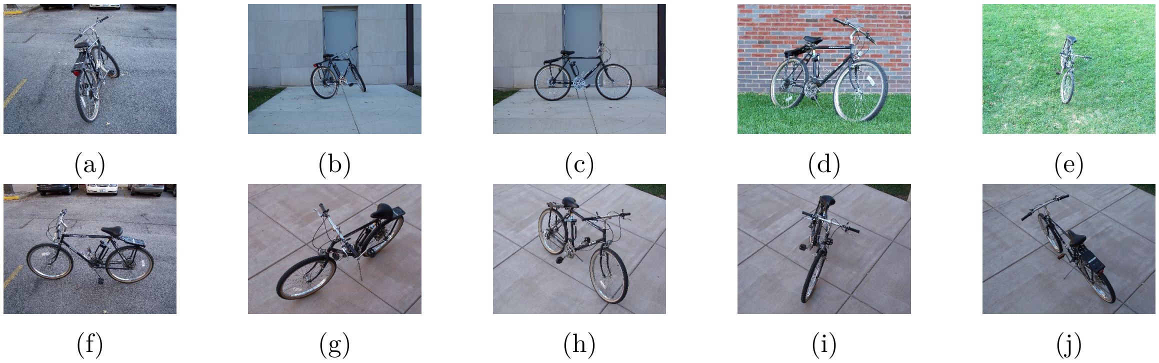


Figure 2.17: Natural scene image Features extraction



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

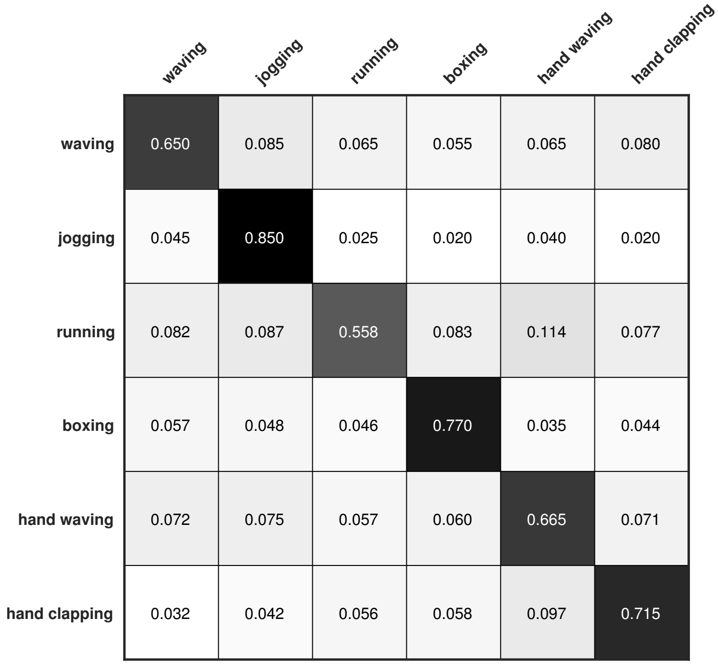
Figure 2.18: An object from a bicycle's class at different 2D views for a 3D modeling

slow as we increase the number of neurons. In this experiment, the best results are obtained with the natural scene dataset (including the face expression data) and the video dataset. One of the challenges when implementing a BPNN is the number of hidden layers needed along with their size.

In overall, the image (natural scene and face expressions) and video datasets provided the best averaged accuracy rates among these 3 classifiers: 68.4%, 64.8%, and 67.82%, respectively at $L = 6$. However, these values are still low compared to the CVB-LGDA's performances on these datasets. This is due the flexibility of its prior which has a very general covariance structure compared to the traditional Dirichlet with very restricted covariance structure. In addition, SVM, KNN, and BPNN are very slow compared to the majority of other methods. It is also justified by their time complexity in Table 2.2. This suggests that for extremely large datasets if speed is your point of interest, these classifiers might not be the best choices. They also tend to use a lot of memory spaces from the time-memory tradeoff concept. This concept is defined as follows: the time complexity is inversely proportional to the space(memory) complexity.

$$TC \propto \frac{1}{MC} \tag{65}$$

TC is the time complexity and MC the memory or space complexity. In a future we will deeply investigate on the distance metrics for the KNN and the Kernels for the SVM as they tend to control or affect the output of the classifier. Learning a proper distance



	waving	jogging	running	boxing	hand waving	hand clapping
waving	0.650	0.085	0.065	0.055	0.065	0.080
jogging	0.045	0.850	0.025	0.020	0.040	0.020
running	0.082	0.087	0.558	0.083	0.114	0.077
boxing	0.057	0.048	0.046	0.770	0.035	0.044
hand waving	0.072	0.075	0.057	0.060	0.665	0.071
hand clapping	0.032	0.042	0.056	0.058	0.097	0.715

Figure 2.19: Confusion matrix of the action classes in video

BPNN		SVM			KNN		
L=4	L=6	A=0.1	A=1	y=10	K=1	K=7	K=10
47.3%	68.4%	57.3%	66.4%	61.8%	48.4%	63.14%	61.22%

Table 2.5: Performance of BPNN SVM and KNN using images (natural scene)

BPNN		SVM			KNN		
L=4	L=6	A=0.1	A=1	A=10	K=1	K=7	K=10
45.9%	64.8%	48.3%	71.25%	47.21%	66.1%	69.4%	58.6%

Table 2.6: Performance of BPNN SVM and KNN using Face expressions

metric for histograms data always plays a central role in computer vision tasks. For our model, another extension could be to equip the clustering algorithm (for the implementation of the codebook) with another distance metric such as the Mahalanobis distance before constructing the Bow features. Basically, we will just remove the Euclidean distances and replace it with the Mahalanobis distance.

Concerning our BoW-based model being evaluated in performance with Generative Adversarial network, we think for now our model could not compete yet with a very sophisticated hybrid deep learning generative model that combines both a generative and discriminative models. For a fair comparison, we are planning to first equip our generative model with a discriminative structure to allow it to be comparable to the GANs models.

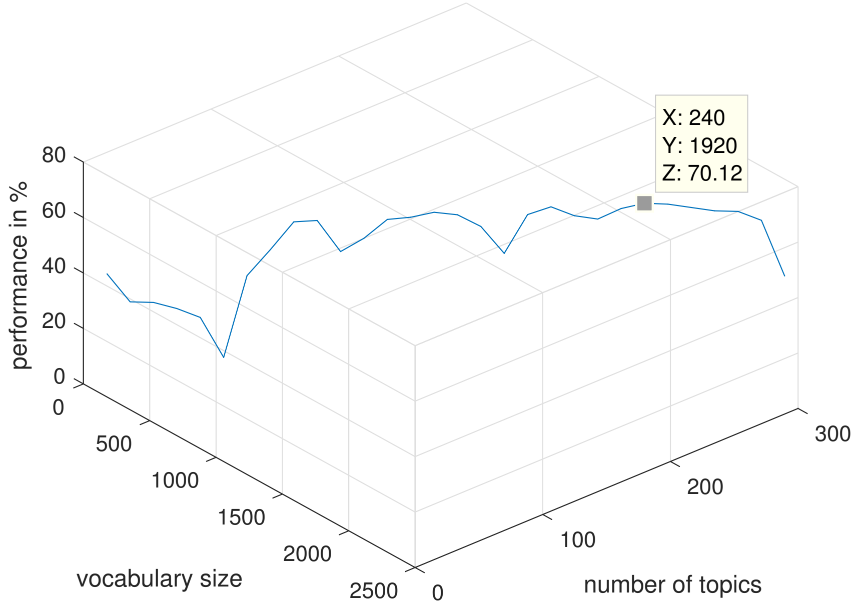


Figure 2.20: Model selection for actions using videos

BPNN		SVM			KNN		
L=4	L=6	A=0.1	A=1	A=10	K=1	K=7	K=10
44.3%	58.2%	58%	60.4%	51.21%	58.1%	59.1 %	58.4%

Table 2.7: Performance of BPNN SVM and KNN using 3D objects

BPNN		SVM			KNN		
L=4	L=6	A=0.1	A=1	A=10	K=1	K=7	K=10
45.73%	67.82%	54.7%	68.1%	61.7%	65.3%	66.1%	54.4%

Table 2.8: Performance of BPNN SVM and KNN using action recognition datasets

2.5 Conclusion

In this chapter, we proposed and implemented a new approach to improve the original LDA hierarchical model. The objective was to provide a strong generalization of the LDA model so that it successfully performs on a variety of datasets besides the usual text data. For this purpose, the new method introduces a flexible GD prior for a robust, complete probabilistic and generative process while maintaining an effective inference technique (CVB). Consequently, the new scheme, the CVB-LGDA is an extension to the GD-LDA, LGDA, and the CVB-LDA. In general, these previous extensions do suffer from two major limitations: incomplete generative processes including the use of priors with very limited capabilities (Dirichlet distribution with very restricted covariance structure) and inefficient inference techniques to build an effective model that could have ability to take into account or handle datasets of different types. Many previous models, were still using the traditional inferences such as VB and CGS (MCMC). These inference schemes have their drawbacks: for instance, the VB suffers from a large bias due to its strong independency assumption between latent variables and parameters. The CGS has a convergence problem. The CVB-LGDA provides a solution to all these different challenges and shortcomings. In the generative process, the new model replaced the Dirichlet distribution on both the corpus and the document parameter with the GD prior, which is shown to be more flexible than the Dirichlet distribution. Doing so, it improved the CVB-LDA, GD-LDA, and the LGDA models. In addition, as consequence of the choice of the GD prior, the CVB-LGDA inference technique is robust, and could perform well in topic correlated environments. Due to the advantages of the GD in topic correlation, the new approach has ability to access a model selection with an optimal number of topics including an optimal vocabulary size (by pruning both irrelevant topics and vocabulary codewords). The amount of correlation between classes (categories) in our experimental datasets showed the flexibilities of the GD prior. It also demonstrates how a positively correlated dataset could hinder the performance in Dirichlet-based LDA models while it is not an issue for the GD-based approaches with the flexibility of the prior's covariance structure.

The performance of the new approach using images, 3D objects, facial expressions, and actions in videos datasets shows the efficiency in the new model. Despite its easy convergence, the CVB-LGDA could be sometimes computationally expensive as it deals with extremely large and complex features from its various descriptors algorithms. The feature extraction could carry a lot of noise that can jeopardize performance if care is not taken in the preprocessing stage. This situation occurred in our images and especially during the 3D and video datasets modeling as some of 2D views of 3D objects were highly corrupted with background noise. Nevertheless, the model was able to provide very satisfactory accuracy rates despite the complexity in these large collections. In addition the new model outperformed its competitors in classification such as the KNN, SVM, and BPNN. As the GAN is getting a lot of attention these days in data analytics with its amazing results, we are planning to equip the CVB-LGDA with a discriminative model to have a fair comparison between the GAN (deep learning) and our model in a categorization framework.

For future work, we will also continue to investigate on the best methods to efficiently perform a preprocessing technique where corrupted background noise effects could be minimized in these datasets. Richer codewords and hierarchies are key to a better performance and result. In addition, we can investigate on other flexible priors to improve our performance in the topic modeling. The model could also be executed in an online fashion to cope with situations where new documents could recursively update the codeword distributions in the database.

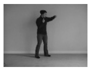
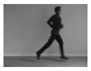
boxing	hand clapping	jogging	hand waving	Running	walking
					

Table 2.9: KTH Action Recognition Dataset

Chapter 3

Stochastic Topic Models for Large Scale and Nonstationary Data

Many traditional database's processing schemes are batch-based with their abilities to utilize the entire information available at a time. Though, their limitations include storage (memory issues) and computational speed (often slow) for large scale applications. Another major disadvantage of the batch processing is that any small change or update in the database often requires a reevaluation using all the data at a time. This is not efficient as it is time consuming and exhausting. So, the approach seems to be a little obsolete in this new generation of fast computation. Furthermore and recently, the decrease in the cost of performing computations online promoted the increase in streaming and online-based models. In other words, new systems are taking advantage of the online setting to build models that are able to perform in real time and handle fast computations with real time updates. Traditional models could no longer scale to very large applications. So, much support has been given to online framework as these massive and nonstationary data could not keep up with the available storage. In the case of generative models, usually, the lack of flexible priors and sometimes the high complexities in the methods often hindered their performances. In addition and most importantly, many online-based models still use traditional inference approaches such as variational Bayes (VB) and Markov chain Monte Carlo (MCMC) which individually are not flexible enough as they suffer from either accuracy or efficiency. As a result, we propose in this chapter, a new model that operates in online fashion with BL (Beta-Liouville) prior due to its flexibilities in topic correlation analysis. Carrying only very few parameters (compared to the generalized Dirichlet distribution, for instance), the BL is now coupled with a robust and stochastic generative process within a new hybrid inference that combines only the advantages of the VB and Gibbs sampling in the collapsed space. This insures an efficient, fast, and accurate processing. Experimental results with nonstationary datasets for face detection, image classification, and text documents processing show the merits of the new stochastic approach.

3.1 Introduction

Owing to internet technology, this era of information could be mainly described by the rapid development of social media platforms where individuals now share knowledge and expertise. Therefore, these sites have not only become repositories of valuable information, but also valuable assets for data science for analytics. Learning from data to minimize costs and risks while improving productivity remains the challenging task for the data scientist. The ability to provide ultimate solutions to industries to maximize profits requires sometimes major changes in the traditional systems and processing methods.

Years ago, batch processing dominated data analytics where high volume of documents and files were stored and processed all at the same time. In other words, in batch processing systems, the entire data is generally presented to the computer's programs that handle them all at once. Ultimately, the scheme has been originally favored when response and feedback (updates) are not needed immediately, and so, batch-based systems can operate for a long period of time, especially unattended. They can therefore deal with large files such as payrolls, bank transactions, billing cycles, examination records, etc. Therefore, batch processing owes its success to batch-based models that continue to support the framework with their ability to perform on large files. The use of generative topic models have also contributed to the extension of batch-based models. Widely used in natural language processing and machine learning, a topic model is a generative technique that has ability to compress a collection of documents into a set of abstract topics or themes [3, 2]. The discovery of these hidden topics (latent variables) in the collections follows the BoW(bag of words) framework.

The Latent Dirichlet Allocation (LDA) in [3] is the first topic model to organize a document into topics using a codebook (dictionary or a vocabulary) from its corpus. Due to the limitations of the Dirichlet prior in text modeling using LDA, data scientists are extending the capabilities of the model in a variety of applications especially in computer vision. This ultimately requires the use of better and more flexible priors such as the BL [58] to improve estimations. Since it has less parameters than the generalized Dirichlet [24, 18, 23], this flexible and conjugate prior to the multinomial has been the distribution of choice in many topic models [50, 35]. Though, one major drawback in these models has been the use of the traditional variational Bayesian inference due to the strong independency assumption in the scheme that tends to affect accuracy in estimation of the likelihood as the lower bound could become unstable. Another important limitation with the batch processing is the constant re-evaluation of the entire system's information in a case of any small change or update in the database. That is time consuming, and as a result not efficient. This also demonstrates the difficulty and complexity in handling theses data as some change with time, and they can become unpredictable. Therefore, the explosion of digital information, the proliferation of large scale datasets have provided a new way of exploring and analyzing data as performances in batch schemes are seriously getting affected by the load, computational speed (slow), memory space (shortage), and convergence problems in the algorithms. In addition, recursive updates are now crucial in order to perform in real time and handle fast computations effectively while managing and optimizing the database's storage.

Then recently, with the invasion of nonstationary data, the online or stochastic processing has started to get a lot of attention in the research community as it can handle one data at a time and exhibit real time updates. This obviously makes the technique fully inline with the new generation of fast computational systems that could provide immediate feedback or update. The huge interest in online systems has also been stimulated by the decrease in the cost of hardware such as supercomputers with their powerful and fast graphics cards: it is very cheap now to perform online processing than it was in the past. Similar to batch systems, online framework has also allowed the implementation of generative models such as [1, 13, 59, 60, 61, 62] that operate in online fashion. The stochastic nature in these methods provides effective data management and storage. Many authors have presented with success the important steps and foundations for online inferences and stochastic learning [13, 63] in generative models ([59, 64]). For instance, the authors in [65, 66] propose a framework that uses latent variables models learning in online fashion while the authors in [4] encourage the collapsed representation for an effective stochastic learning. As a result, considering the major contributions from previous models and frameworks, our proposed approach ultimately implements an online inference technique that has ability to handle a variety of applications other than the traditional text document analysis often observed in the previous schemes. In other words, our developed framework mainly emphasizes on images and videos processing (computer vision) with also a focus on the use of a better prior distribution for efficiency and flexibility in the generative probabilistic models using the BoW approach.

However, many online schemes were built with the VB (variational Bayesian inference), making them limited in terms of performance [12]. While some authors using nonparametric settings implement models that have too much complexities for an already computationally expensive approach [59, 62], the traditional Gibbs sampling method is not a deterministic technique [12]. As a result, we propose a new model, the stochastic collapsed variational Bayesian-based latent Beta-Liouville allocation (SCVB-LBLA) which is implemented in the collapsed space of latent variables. As a hybrid model, it integrates the advantages of VB and CGS (Collapsed Gibbs Sampler) as in [12] where it now uses the BL in the generative process in online fashion. Experimental results with nonstationary data in image recognition and categorization, videos, and text analysis show the merits of the new approach which carries the online framework using the minibatch scheme. Its contribution can be summarized as follows:

- While using the robust minibatch method, the new model with the BL prior has few parameters to estimate, so its MLE (maximum likelihood estimation) performs faster than the case of generalized Dirichlet-based models ([18, 64, 67]). It therefore has a better time complexity compared to its major competitors, and the stochastic nature of the model could handle nonstationary data better than batch approaches. This ultimately improves computational speed and data storage.

- It is implemented with an efficient inference technique using a prior distribution that could effectively facilitate online topic correlation learning and vocabulary analysis. As a result, we can reach a model selection scheme that considers both the optimal number of topics and the size of the codebook (vocabulary).

- This approach carries the advantages of the collapsed representation that is known to be suitable for stochastic inference. Its collapsed variational Bayesian (CVB) inference is an improved version of the variational Bayes since the collapsed representation provides a better lower bound that is stable for the parameters estimation. The model is so flexible it could be used in a wide variety of applications such as retrieval, classification, recognition, and analysis.

- It can ultimately handle faster large scale corpora's processing with efficiency, due to its memoryless (online) structure as there is no need to maintain local estimates (distributions) during update as in [13, 65].

This chapter is structured as follows: section 4.2 illustrates the background and related work. Section 3.3 presents the new approach (which introduces a batch model before the online scheme) while section 4.4 covers the experiments and results in several applications. Finally, section 4.5 explores some future work and provides a conclusion.

3.2 Related Work And Background

The original batch-based topic model was implemented using the LDA (latent Dirichlet allocation) model. As a generative probabilistic model, the LDA uses the BoW framework where its documents are usually represented as frequency counts. Its generative process is widely detailed in so many publications such as [3, 2, 13]. Based on this complete generative process, the one for the LBLA [35] using K topics is easily defined as follows:

```
Generate each topic $\varphi_k \sim$ Beta-Liouville($\zeta$) where $k \in 1, ..., K$
For each document $j$
    Generate a distribution over topics $\theta_j \sim$ Beta-Liouville ($\varepsilon$)
      For each word $i$ in document $j$
          Sample a topic $z_{ij} \sim$ Discrete($\theta_j$) or Multinomial($\theta_j$)
          Sample the word $w_{ij} \sim$ Discrete($\varphi_{z_{ij}}$) or Multinomial($\varphi_{z_{ij}}$)
```

It is noteworthy to mention that in a case of a classification, the model generates the class, then the topics followed by the words for the class using the framework proposed in [2] for the LDA.

The batch-based models were widely used because of their ability to process all the information available at a time. Due to their success, many authors have provided important contributions for the batch processing. For instance, it has been proved in [13] that concerning the batch modeling, inference schemes that usually operate in the collapsed space (where parameters are marginalized out leaving only the latent variables) can improve probabilities of the held-out compared to the uncollapsed space also known as the joint space of latent variables and parameters. Moreover, this collapsed representation provides a better variational bound as it induces models with fewer parameters to update where also the digamma functions estimations are finally avoided. For the records, the digamma functions are known for slowing down computations and updates. As a result, despite the use of flexible priors such as the BL, batch-based models including inference schemes such as [50, 35] are still not robust enough to provide good performance since they all operate in the uncollapsed space within the VB inference. This could affect the accuracy in predictive models [13]. The variational Bayes alone as an inference technique could have a limitation due to the strong independency assumption between the latent variables and the parameters. This situation could negatively have an impact on the lower bound and the model likelihood probability function as they could become loose and inaccurate [12].

Despite the tremendous success in the field of topic modeling, it is noteworthy that one of the major disadvantages in batch processing remains the fact that it is usually time consuming as a small change or update in the system database often require a reevaluation of the entire database. Consequently, extensions to the state-of-the-art batch-based approaches such as the CVB0 [4] and the CVB-LGDA (collapsed variational Bayesian inference technique for latent generalized Dirichlet allocation) [67] are still limited in terms of flexibility in the performance. In fact, the drawback in the CVB0 for instance includes the extremely large memory requirement it often needs to store a variational distribution over the topic assignment for every word in the corpus [4]. In general, these previous techniques and extensions to the LDA [3] are not capable of scaling to very large datasets including nonstationary data [13]. As a result, concerning batch processing, these shortcomings have finally made online schemes suitable for fast and efficient computional systems. However, many of the current existing online-based inference techniques in topic modeling also do not fully take advantage of the collapsed space either as the majority of them such as [1, 13, 59, 60, 61, 62] are still variational-based within the uncollapsed space. These models are often implemented with priors with limitations such as the Dirichlet or built within a nonparametric setting (Dirichlet processes) that ultimately increases the model computational complexity. One of the first collapsed representations is the sparse online LDA proposed in [68]. As based on Dirichlet, this online model in the collapsed space only marginalizes the documents parameters while leaving the corpus parameters. Shortly, the memory issues in the CVB0 ultimately leads to the stochastic CVB0 or the SCVB0 [13].

While maintaining all the advantages of the collapsed representation, the SCVB0 also solves the memory space problem in the batch processing in the CVB0. The CVB0 has been previously known as the fastest method for single-core batch inference due to its convergence rate [4, 13]. Now, the SCVB0 algorithm does no longer need to maintain the variational distribution on every word, solving therefore the memory issue in the LDA. And, despite the success of this online-based model for text documents, the technique is still Dirichlet-based (as the batch models in [3, 69]), and so, could not be effective in topic correlation [18] and vocabulary analysis [67]. The constant problem with the Dirichlet prior is its very restrictive covariance structure (negative correlation) which can hinder performance for modeling positively correlated datasets [7].

Considering all these challenges and shortcomings, we propose in this chapter, the SCVB-LBLA (stochastic collapsed variational Bayesian inference for the latent Beta-Liouville allocation) model. It is a direct extension to the SCVB-LDA (stochastic collapsed variational Bayesian inference for the latent Dirichlet allocation)[4] which so far operates on text documents only. One of the goals here is to extend the capabilities of the LDA structure and its generative scheme using for instance nonstationary data (in online fashion). Consequently, we emphasize on experimental results using applications related to a wide variety of nonstationary datasets that include images, videos, and texts to show the flexibilities and merits of the new approach compared to its major competitors such as [67, 1, 13, 59, 60, 61, 62]. The new generative scheme in our method, in contrast to the one proposed in [67], has now the corpus and documents parameters all drawn from the BL instead of using the generalized Dirichlet (GD). In addition, as the new technique is an extension to the batch-based LBLA [35], we will first present the CVB-LBLA (collapsed variational Bayesian inference for the latent Beta-Liouville allocation) model before the stochastic SCVB-LBLA which is ultimately the online version of the CVB-LBLA. It is noteworthy that in the LBLA proposed in [35], the document parameter is drawn from a BL while the corpus parameter remains non generated to facilitate parameter estimation during MLE. Such approach is acceptable in VB. However, the collapsed representation using the Gibbs sampler, as in our proposed approach, requires both the corpus and the document parameters to be sampled (in this case from BL). As the authors [1, 66] provided the platform for online topic modeling, we are taking advantage of their framework for a robust extension of the LDA using the BL prior.

3.2.1 Beta-Liouville distribution

The LDA and its Dirichlet prior have a very rich literature in text modeling and computer vision. The generative process in LDA has been extensively described in so many publications [3, 69, 67]. Often presented as a solution to the limitation of the Dirichlet distribution, the Beta-Liouville (BL) distribution is a generalization of the Dirichlet prior. In other words, the Dirichlet could be seen as a special case of the BL. In fact, the BL has a more general covariance structure. This shows the flexibility of the BL for its use in a variety of data such as non Gaussian proportional data (normalized histograms) [11, 59] and count data [67].

Mathematically, these two conjugate priors (Dirichlet and BL) to the multinomial distributions could be defined as follows: In a $K-$dimensional space, the $K-$dimensional random variable (vector) $\vec{\theta}$ following a Dirichlet distribution with hyperparameters ε could be expressed as:

$$p(\vec{\theta}|\varepsilon) = \frac{\Gamma(\sum_{k=1}^{K} \varepsilon_k)}{\prod_{k=1}^{K} \Gamma(\varepsilon_k)} \prod_{k=1}^{K} \theta_k^{\varepsilon_k - 1} \tag{66}$$

such that $\sum_{k=1}^{K} \theta_k = 1$. Though, in a $(K+1)-$dimensional space, any $K-$dimensional vector $\vec{\pi}$ drawn from the Beta-Liouville distribution with hyperparameters $\varrho = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ is defined as:

$$p(\vec{\pi}|\varrho) = \frac{\Gamma\left(\sum_{d=1}^{K} \alpha_d\right) \Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha\right) \Gamma\left(\beta\right)} \prod_{d=1}^{K} \frac{\pi_d^{\alpha_d - 1}}{\Gamma\left(\alpha_d\right)} \left(\sum_{d=1}^{K} \pi_d\right)^{\alpha_d - \sum_{d=1}^{K} \alpha_d} \times \left(1 - \sum_{d=1}^{K} \pi_d\right)^{\beta - 1} \tag{67}$$

If $\vec{\pi} = (\pi_1, ..., \pi_K)$ follows a Beta-Liouville distribution with hyperparameter ϱ and the vector of counts $\vec{X_i} = (X_1, ..., X_{D+1})$ follows a multinomial distribution with parameter $\vec{\pi}$, then the posterior distribution $p(\vec{\pi}|\varrho, \vec{X_i})$ is also a Beta-Liouville [35] due to the conjugacy property between the BL prior and the multinomial distribution where the parameter updates are given as :

$\alpha'_d = \alpha_d + X_d \qquad \alpha' = \alpha + \sum_{d=1}^{D} X_d \qquad \beta' = \beta + X_{D+1}$

3.2.2 Time complexity

As data increase in size and complexity, this does not translate automatically into a better performance in traditional models because they cannot scale to very large data. It means these models could not accurately and efficiently operate on high volume of datasets. For instance, in batch learning, for a corpus of N words, within the BoW, and K topics, the time complexity for a batch-based model is about $O(NK)$ in LDA with the VB inference. However, this time complexity is a little smaller with the CGS [12]. With the CVB inference scheme, the computational cost scales up to $O(MK)$ where M is the number of unique words in the corpus. So we can observe that the computation is faster in CVB-LDA compared to the LDA as $N > M$. In CVB-LGDA [67], the time complexity is almost close to the one in CVB-LDA, but the model is able to perform more tasks such as semantic relationship between words, topic correlation, and vocabulary analysis. It makes the CVB-LGDA more flexible in batch processing than the CVB-LDA and the LDA, as it implicitly performs faster each one of those tasks.

In the stochastic framework with minibatches, let's assume Υ being a minibatch and $|\Upsilon|$ the size of each minibatch to follow the definition in [13]. The time complexity will be just about $O(|\Upsilon|K)$. However, usually in the online scheme, the model does not need to analyze all the documents before updating its parameters. It makes sense the stochastic learning is very fast as its time complexity could be a little less than $O(|\Upsilon|K)$ since $|\Upsilon|$ could be very small. Therefore, we can make a very quick conclusion that the SCVB-LBLA and even the SCVB-LDA have arguably the same time complexity of $O(|\Upsilon|K)$. Nevertheless, the SCVB-LBLA is more complex and far superior compared to the SCVB-LDA. In other words, the SCVB-LBLA, as previously mentioned, has also ability to execute many tasks at once including topic correlation analysis, vocabulary analysis if needed, and semantic analysis between words. It implies that it just performs each one of these tasks faster than the SCVB-LDA which is limited as it could not even process positively correlated data due to the Dirichlet prior. The time complexity also justifies the reason the online scheme is preferred compared to the batch methods. The batch techniques are usually slow and require more memory space while the stochastic learning models are often faster with a low memory requirement, ideal framework for large scale applications.

3.3 The New Approach

The original LBLA (latent Beta-Liouville allocation model) in [35] manages to replace the Dirichlet prior in the original LDA with the BL while using the VB inference. We are providing an extension (SCVB-LBLA) to this work in online fashion where we implement instead the collapsed variational Bayesian inference which is a hybrid inference between the VB and CGS also using the BL prior distributions. Though, we will first introduce the batch-based CVB-LBLA model since such setting will allow the readers to better understand the motivations around our proposed online version.

3.3.1 Overview: description, notations, and definitions within the bag of word framework

In this chapter, we propose a model that uses the Beta-Liouville (BL) prior on both the document and corpus parameters in a collapsed space for its flexibility [23, 36]. The BL has a more general and versatile covariance structure than the Dirichlet prior. It also has less parameters compared to the GD [23]. A variational inference scheme with this conjugate prior in the collapsed space is an improvement to the state-of-the-art topic modeling inference proposed in [12] within the BoW framework. In addition, the new model deals with challenges related to an extensive vocabulary size, and increasing number of topics. The approach integrates two models: a topic model (unsupervised learning) and a classification model (supervised learning).

As we are planning to implement inferences in the two spaces (the collapsed space and the joint space), details about each of them will be provided along with their characteritics and variables that define each one of them. Though, to briefly describe the architecture in our model which is based on a graphical representation similar to the smoothed LDA proposed in [3], the variable ε carries the document hyperparameters α and β while ζ holds the corpus hyperparameters η and λ. Moreover, the couple (ε,c) is defined as the hyperparameter set for a document in a class c where it is extended as: $(\varepsilon, c) = (\alpha_{c1}, ..., \alpha_{cK}, \alpha_c, \beta_c)$ with K as the number of topics in the corpus. Similarly, the hyperparameter variable ζ can also be extended as: $\zeta = (\lambda_1, ..., \lambda_V, \lambda, \eta)$ where the subscript V is the size of the codebook. In our method, documents are drawn from the class set c, where their parameters θ and the corpus parameters φ are sampled from the Beta-Liouville distributions. In implementation, the hyperparameter variable ε holds two $1 \times C$ vectors α_c and β_c and a single $C \times K$ matrix

Table 3.1: Model variables and definitions

Model Variables	
D	Total number of documents
N	Total number of words in a document
K	Total number of topics
V	Vocabulary size
(i, j)	ith word or topic assignment in the jth document
k	kth topic
$x = \{x_{ij}\}$	Observed words
$Z = \{z_{ij}\}$	Latent variables
$\theta_j = \{\theta_{jk}\}$	Mixing proportions
$\varphi_k = \{\varphi_{kw}\}$	Corpus parameters
$BL(\varepsilon)$	Beta-Liouville distribution with parameter ε
$\theta_{jk}/\varepsilon \sim BL(\varepsilon)$	θ_{jk}/ε drawn from $BL(\varepsilon)$
$\varphi_{kw}/\zeta \sim BL(\zeta)$	φ_{kw}/ζ drawn from $BL(\zeta)$
$Mult(\theta_{jk})$	Multinomial distribution with parameter (θ_{jk})
$z_{jk}/\theta_{jk} \sim Mult(\theta_{jk})$	z_{jk}/θ_{jk} drawn from $Mult(\theta_{jk})$
$x_{jk}/z_{jk}, \varphi_{jk} \sim Mult(\varphi_{kw})$	$x_{jk}/z_{jk}, \varphi_{jk}$ drawn from $Mult(\varphi_{kw})$
$c = \{1, 2, ..., C\}$	Number of classes
N^{-ij}	Counts where the superscript $-ij$ denotes the corresponding variables with x_{ij} and z_{ij} excluded
Υ	A minibatch

where each row is the K-dimensional vector $(\alpha_{c1}, ..., \alpha_{cK})$ such that ε_c is K-dimensional BL hyperparameter $(\alpha_{c1}, ..., \alpha_{cK}, \alpha_c, \beta_c)$ for the document in a class c in a $(K+1)$-dimensional space. Similarly, for every topic k, the hyperparameter variable ζ contains one vector of size $V \times 1$, $(\lambda_1, ..., \lambda_V)$ and two strictly positive constants (λ and η) such that ζ is a V-dimensional BL hyperparameter $(\lambda_1, ..., \lambda_V, \lambda, \eta)$ for the corpus in a $(V+1)$-dimensional space.

Furthermore, in inferences, the new approach uses concepts of variational distributions and variational lower bound to help approximating the posterior distributions. Therefore, following the work in [12], the variable $\tilde{\Delta}$ is the variational distribution in the standard space or the uncollapsed space (the joint space of parameters and latent variables or the uncollapsed space [13]). However, $\hat{\Delta}$ is the variational distribution of the collapsed space of latent variables where the parameters are marginalized out. In LDA, using the exponential family distributions, the likelihood function (the normalization factor in the posterior distribution) is often approximated by a lower bound defined as $\exp(\mathcal{F}(\Delta(x)))$, where $\mathcal{F}(\Delta(x))$ is the variational lower bound in the log space [37]. This functional is called the variational free energy [12]. One objective is to demonstrate that our new model is an improved stochastic variational Bayes scheme in the collapsed space of latent variables compared to the traditional batch-based VB inference that is performed in the joint space of latent variables and model parameters. The traditional VB is slow when compared to the one in the collapsed space. We can finally define the bounds to show all the steps for the implementation of the new approach in the collapsed space. As a result, similar to the variational distributions $\tilde{\Delta}$ and $\hat{\Delta}$, the variable $\tilde{\mathcal{F}}$ is the variational bound in the joint (uncollapsed) space, and $\hat{\mathcal{F}}$ is the variational bound in the collapsed space.

Widely used in computer vision, the bag of word (BoW) framework can represent any data as a collection of documents containing frequency counts, after implementation of the model codebook. Using the bag of word (or visual word) architecture, an image patch can be assimilated to a word, and it is the basic unit in a document while a document $\mathcal{X}$ itself is a collection of N patches (words) such that $\mathcal{X} = (x_1, x_2, ..., x_N)$. The variable x_n is the nth patch in the document. A category is a group of D documents within the same class such that $I = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_D\}$.

3.3.2 Batch Learning: inferences

Our proposed approach is composed of two main inferences: the variational Bayes and the collapsed Gibbs sampling. These two methods lead to the collapsed variational Bayesian inference that is a hybrid technique combining the CGS and the VB. The (batch) LBLA model proposed in [35] has proved in various applications to outperform the original LDA. In this chapter, one of the objectives, is still to show that our batch-based CVB-LBLA model also represents an extension to the LBLA and LDA, with a better generative process and a robust inference technique [67].

This hybrid model combines the advantages of the VB and the CGS while maintaining accuracy and efficiency in its hybrid inference. It is noteworthy that once the batch CVB-LBLA framework is implemented, its stochastic version, that is our main model in this chapter, should be easy to understand.

3.3.2.1 General Batch-based Bayesian inference framework in the joint space

In machine learning, the Bayesian inference computes the posterior distribution (of the hidden variables) given the observations. However, the estimation of the posterior in topic modeling literature involves integrals evaluation in the likelihood function that turns out to be intractable. It does finally make the posterior distribution also not tractable due to its direct relationship with the likelihood function in the Bayesian framework. Consequently, inference methods such as VB and MCMC (Gibss sampling) are often implemented to estimate the latent topics and the model parameters. For instance, given the hyperparameters ε, ζ, and the class parameter μ, we can express the full generative equation of the model as a joint probability distribution noted $p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu)$:

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) \prod_{i=1}^{K} p(\varphi_i|\zeta) \prod_{j=1}^{D} p(\theta_j|\varepsilon, c) \prod_{n=1}^{N} p(z_{j,n}|\theta_j) p(x_{j,n}|\varphi z_{j,n}) \tag{68}$$

This joint distribution's equation can be reduced to:

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) p(\theta|c, \varepsilon) p(\varphi|\zeta) \prod_{n=1}^{N} p(z_n|\theta) p(x_n|z_n, \varphi) \tag{69}$$

where $p(\varphi|\zeta)$ and $p(\theta|c, \varepsilon)$ are the corpus BL (prior) distribution with hyperparameters ζ and document BL (prior) distribution with hyperparameter ε in class c, respectively. The distributions $p(z_n|\theta)$ and $p(x_n|\varphi z_n)$ are multinomials while $p(c|\mu)$ is the class prior distribution. The Bayesian inference estimates the joint posterior distribution of the latent variables z and the model parameters (θ and φ) given the observations and the class, $p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu)$ as seen in the equation below:

$$p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu) = \frac{p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu)}{p(\mathcal{X}, c|\varepsilon, \zeta, \mu)} \tag{70}$$

where its denominator is given as :

$$p(\mathcal{X}, c|\varepsilon, \zeta) = \int_\theta \int_\varphi \sum_z p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta) d\varphi d\theta \tag{71}$$

with

$$p(\mathcal{X}, c|\varepsilon, \zeta, \mu) = p(\mathcal{X}|\varepsilon, \zeta, c)p(c|\mu) \tag{72}$$

For a uniform class prior, we obtain $p(c|\mu) = p(c) = \frac{1}{C}$ with μ negligible. As a result, the Eq.71 and Eq.72 could be reduced to:

$$p(\mathcal{X}, c|\varepsilon, \zeta, \mu) = \frac{p(\mathcal{X}|\varepsilon, \zeta, c)}{C} \tag{73}$$

C is the total number of classes while c is the set of classes in this graphical model. The posterior distribution then becomes:

$$p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu) = \frac{p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu)}{p(\mathcal{X}|\varepsilon, \zeta, c)/C} \tag{74}$$

As mentioned previously, the class conditional $p(\mathcal{X}|c, \varepsilon, \zeta)$ also known as the likelihood function is not tractable implicitely making the posterior $p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu)$ also intractable. The variational Bayes (VB) technique apprroximates the true posterior distribution using variational distributions [3, 12] $\tilde{\Delta}(z, \theta, \varphi)$ that could be factorized as:

$$\tilde{\Delta}(z, \theta, \varphi) = \prod_{ij} \tilde{\Delta}(z_{ij}|\tilde{\psi_{ij}}) \prod_j \tilde{\Delta}(\theta_j|\tilde{\varepsilon_j}) \prod_k \tilde{\Delta}(\varphi_k|\tilde{\zeta_k}) \tag{75}$$

where $\tilde{\Delta}(z_{ij}|\tilde{\psi_{ij}})$ is the variational multinomial distribution with parameters $\tilde{\psi_{ij}}$. In addition, $\tilde{\Delta}(\theta_j|\tilde{\varepsilon_j})$ and $\tilde{\Delta}(\varphi_k|\tilde{\zeta_k})$ are the BL variational distributions with parameters $\tilde{\varepsilon_j}$ and $\tilde{\zeta_k}$, respectively, in the joint space of latent variables and model parameters. The standard and traditional VB operates in the joint space of latent variables and parameters, and this inference always requires a set of variational distributions, defined as $\tilde{\Delta}(z, \theta, \varphi)$ that should be close to the true posterior distribution $p(z, \theta, \varphi|c, \varepsilon, \zeta)$ with the KL (KullBack Leibler) divergence. This leads to an optimization scheme [3] as it is defined below:

$$(\tilde{\psi}^*_{ij}, \tilde{\varepsilon}^*_j, \tilde{\zeta}^*_k) = \text{argmin}\, \tilde{\psi}_{ij}, \tilde{\varepsilon}_j, \tilde{\zeta}_k D(\tilde{\Delta}(z, \theta, \varphi|\tilde{\psi}_{ij}, \tilde{\varepsilon}_j, \tilde{\zeta}_k)||p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu)) \tag{76}$$

The variational Bayesian inference always provides a lower bound to the marginal log likelihood function; that is equivalent to the VB upper bounding the negative marginal log likelihood $-\log p(\mathcal{X}|c, \varepsilon, \zeta)$ in a scheme [12] that utilizes the concept of variational free energy (Eqs. 77 and 80). As a deterministic approach, the VB is efficient since it is easy to implement with an easy access to convergence. The inference computes the variational parameters updates and finally the model parameters in an expectation-maximization (EM) method. Using Eqs. 77 to 81, the lower bound on the log likelihood is expressed as:

$$\log p(\mathcal{X}|c, \varepsilon, \zeta) \geq \int_\theta \int_\varphi \sum_z \Delta(z, \theta, \varphi) \times \log p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta) d\varphi d\theta \tag{77}$$

$$- \int_\theta \int_\varphi \sum_z \Delta(z, \theta, \varphi) \log \Delta(z, \theta, \varphi) d\varphi d\theta \tag{78}$$

$$= \mathbb{E}_\Delta[\log p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta)] - \mathbb{E}_\Delta[\log \Delta(z, \theta, \varphi)] \tag{79}$$

$$
\begin{aligned}
-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq & -\int_{\theta}\int_{\varphi}\sum_{z}\Delta(z,\theta,\varphi)\times\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)d\varphi d\theta \\
& +\int_{\theta}\int_{\varphi}\sum_{z}\Delta(z,\theta,\varphi)\log Q(z,\theta,\varphi)d\varphi d\theta \\
& =\mathbb{E}_{\Delta}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)]-\mathbb{E}_{\Delta}[-\log\Delta(z,\theta,\varphi)]
\end{aligned}
\tag{80}
$$

$$
-\log p(\mathcal{X}|c,\varepsilon,\zeta)\leq\tilde{\mathcal{F}}(\tilde{\Delta}(z,\theta,\varphi))=\mathbb{E}_{\tilde{\Delta}}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)]-\mathcal{H}(\tilde{\Delta}(z,\theta,\varphi)) \tag{81}
$$

While the variational entropy $\mathcal{H}(\tilde{\Delta}(z,\theta,\varphi))$ is computed as $\mathcal{H}(\tilde{\Delta}(z,\theta,\varphi)) = \mathbb{E}_{\tilde{\Delta}}[-\log\tilde{\Delta}(z,\theta,\varphi)]$, the variational posterior distribution in the joint space $\tilde{\Delta}(z,\theta,\varphi)$ is factorized using the independency assumption (Eq. 75). In the joint space using the VB with the BL prior, the model follows the EM algorithm framework where its parameters θ, φ are estimated in M-step after an E-step that updates the variational distributions hyperparameters from the estimate variational multinomial parameter $\tilde{\psi}_{ijkc}$. Many publications in topic modeling have covered the implementation of the Dirichlet-based VB inferences [3, 2, 12, 39, 23]; however, the limitations often observed in the Dirichlet prior coupled with the strong independency assumption in VB (that could in overall jeorpardize performances) ultimately led to changes in our proposed Bayesian inferences that aim to improve traditional techniques instead.

The CVB is not only a combination of VB and MCMC approaches, but also an improved version of the VB method in the collapsed space of latent variables [12, 40, 13]. It is the state-of-the-art inference we are upgrading in this chapter with a better and more flexible prior that is the Beta-Liouville distribution mainly for online learning. As the CVB and the CGS both operate in the collapsed space of latent variables, in the joint distribution $p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta,\mu)$, the model parameters θ, φ are marginalized out to obtain the marginal joint distribution $p(\mathcal{X},z,c|\varepsilon,\zeta)$ defined as:

$$
p(\mathcal{X},z,c|\varepsilon,\zeta)=\int_{\theta}\int_{\varphi}p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)d\varphi d\theta \tag{82}
$$

But $p(\mathcal{X},z,c|\varepsilon,\zeta)=p(\mathcal{X},z|c,\varepsilon,\zeta)p(c)$ so $p(\mathcal{X},z|c,\varepsilon,\zeta)$ becomes

$$
p(\mathcal{X},z|c,\varepsilon,\zeta)=C\int_{\theta}\int_{\varphi}p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)d\varphi d\theta \tag{83}
$$

One of the advantages of the collapsed representation is the easiness in computing the integral above (Eq. 83) which becomes a product of Gamma functions (Eq. 93), avoiding therefore the difficulty to perform parameters estimation with digamma functions in the joint space as they tend to slow down processing in VB and updates. This advantage in VB is due to the conjugacy property between the BL and the multinomial distribution. The ultimate goal is to approximate the conditional distribution of the latent variables $p(z|\mathcal{X},c,\varepsilon,\zeta)$ through an efficient sampling process.

3.3.2.2 Variational Bayes with BL prior: BL-based VB

As a deterministic approach and in contrast to the CGS, the VB insures convergence to a local minimum with the EM (Expectation-Maximization) algorithm [3]. Following the method in [12], optimizing the variational distribution in Eq. 75 from Eq. 81 with respect to the BL variational parameters leads to the following updates in the variational distributions:

$$\tilde{\alpha}_{jkc} = \alpha_{ck} + \sum_i \tilde{\psi}_{ijkc} \tag{84}$$

$$\tilde{\alpha}_{jc} = \alpha_c + \sum_i \tilde{\psi}_{ijkc} \tag{85}$$

$$\tilde{\beta}_{jc} = \beta_c + \tilde{\psi}_{ij(K+1)c} \tag{86}$$

$$\tilde{\lambda}_{kw} = \lambda_w + \sum_{ij} \vec{1}(x_{ij} = w)\tilde{\psi}_{ijkc} \tag{87}$$

$$\tilde{\lambda}_w = \lambda + \sum_{ij} \vec{1}(x_{ij} = w)\tilde{\psi}_{ijkc} \tag{88}$$

$$\tilde{\eta}_{kw} = \eta + (x_{ij} = w')\tilde{\psi}_{ijk} \tag{89}$$

where w' is the $(v+1)^{th}$ codeword in the vocabulary or codebook. The multinomial update or count $\tilde{\psi}_{ijkc}$ is also obtained through an optimization of the joint posterior variational distribution $\tilde{\mathcal{F}}(\tilde{\Delta}(z))$ with respect to the multinomial variational parameter [12].

From observations, these updates look similar to those found in the LBLA proposed in [35]. Nevertheless, the difference is noticeable as we can see here the updates in the corpus (Eqs. 87 to 89). In the LBLA, only the document parameters were updated. This provides another flexibility in our new model that aims to show efficiency and accuracy in inferences. In the joint space, the BL variational parameter $\tilde{\alpha}_{jkc}$ is a document-topic count as it is the total number of words in a topic k in a document j, all in a class c. The BL variational parameter $\tilde{\alpha}_j$ is also a document-topic count defined as the total number of words in the first K topics in a document j. Finally, $\tilde{\beta}_j$ is a document-topic count for the $(K+1)^{th}$ topic. Similarly, in the corpus, the BL variational parameter $\tilde{\lambda}_{kw}$ is defined as a word-topic count, and it is the number of times a word w from a codebook appears in a topic k while $\tilde{\lambda}_{kw}$ is another word-topic count that suggests the total number of times the first V codebook words appear in the corpus. The variational parameter $\tilde{\eta}_{kw}$ is the number of times the $(V+1)^{th}$ word appears in the corpus. The VB has a well defined convergence criterion [12, 3, 5], but often the inference techniques suffers from a large bias due to its strong independency assumption which allows to decouple the joint variational posterior into a product of individual and independent variational posterior distributions. This is because the scheme always assumes (for convenience) that latent variables and model parameters are independent in the true posterior distribution. In the case this assumption fails, this could make the current VB inferences inaccurate as the lower bound in this case could no longer be stable, affecting therefore the log likelihood function and the possibility for a true mean field approximation [12].

3.3.3 The Collapsed space: CGS, CVB, and CVB0

It is the space where the parameters are marginalized out leaving only the latent variables that become conditionally independent given the parameters [12]. We will later show that the collapsed representation is suitable for models that operate in online fashion [70]. The collpased space inferences are often represented by the CGS, the CVB, and the CVB0 which is a simple version of the CVB [13]. Performing estimation in a reduced space is necessary for fast computation. And the collapsed representation offers that property and advantage as parameters are marginalized out. In addition, the BL has few hyperparameters than the GD (generalized Dirichlet) distribution which implies that this model will be fast when performing MLE.

3.3.3.1 The Collapsed Gibbs sampler and Mean field inference

The low dimensional space provided by the collapsed representation with its reduced number of hidden variables (as parameters are integrated out) offers the possibility for true mean field inference. In addition, as mentioned previously, it is suitable for easy computation of integrals. In this space, the collapsed Gibbs sampler provides inference by computing expectations through a sampling process of the latent variables. This aims to approximate the posterior distributions using a network of conditional probabilities (Bayesian network). The CGS [12, 18, 41] in the collapsed space of latent variables allows a very fast estimation compared to the standard Gibbs sampler that operates in the joint space of latent variables and model parameters. Obviously, with the CGS, no more use of digamma functions which are computationally very expensive in VB method and updates. The CGS algorithm is required to estimate the parameters when the Markov chain reaches its stationary state (stationary distribution) where it provides the most accurate estimate of the true posterior distribution than the VB.

The marginal joint distribution $p(\mathcal{X}, z|c, \varepsilon, \zeta)$ in Eq. 83 is in integral form. Using this integral, the conditional probabilities of the latent variables z_{ij} are estimated given the current state of all variables while discarding the single variable z_{ij}[12]. The collapsed Gibbs sampler estimates the topic assignments associated with the observed words using the conditional probability of latent variables $p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ where $-ij$ corresponds to counts or variables with z_{ij} discarded [12]. This conditional probability is defined as:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) = \frac{p(z_{ij}, z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)}{p(z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)} \tag{90}$$

Following the work in [12], it can still be simplified since:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) \propto p(z_{ij} = k, z^{-ij}, \mathcal{X}, c|\varepsilon, \zeta) \tag{91}$$

The obtained Callen equations (below) as in [12] demonstrate the way the collapsed Gibbs performs its sampling mechanism that can be finally summarized as an expectation problem:

$$p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta) = \mathbb{E}_{p(z^{-ij}|c,\mathcal{X},\varepsilon,\zeta)}[p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)] \tag{92}$$

3.3.3.2 The New collapsed space with the BL prior

Following the work in [12], it is faster to sample in the collapsed space of just latent variables than it is in the joint space of both latent variables and parameters [12]. The motivation here is to sample the latent variables from the joint distribution $p(\mathcal{X}, z|c, \varepsilon, \zeta)$ using a network of single class conditional probabilities. The conjugacy assumption facilitates estimation of the integral in Eq. 83 obtained as a product of Gamma functions (Eq. 93).

$$p(\mathcal{X}, z|c, \varepsilon, \zeta) = C \prod_{j=1}^{M} \left[\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_{ci}\right) \Gamma\left(\alpha_c + \beta_c\right)}{\Gamma\left(\alpha_c\right) \Gamma\left(\beta_c\right) \prod_{i=1}^{K} \Gamma\left(\alpha_{ci}\right)} \right] \times \frac{\Gamma\left(\alpha'\right) \Gamma\left(\beta'\right) \prod_{i=1}^{K} \Gamma\left(\alpha'_{ci}\right)}{\Gamma\left(\sum_{i=1}^{K} \alpha'_{ci}\right) \Gamma\left(\alpha' + \beta'\right)} \times \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \lambda_r\right) \Gamma\left(\lambda + \eta\right)}{\Gamma\left(\lambda\right) \Gamma\left(\eta\right) \prod_{r=1}^{V} \Gamma\left(\lambda_r\right)} \times \frac{\Gamma\left(\lambda'\right) \Gamma\left(\eta'\right) \prod_{r=1}^{V} \Gamma\left(\lambda'_r\right)}{\Gamma\left(\sum_{r=1}^{V} \lambda'_r\right) \Gamma\left(\lambda' + \eta'\right)} \tag{93}$$

where the document-topic update in a class is expressed as:

$$\begin{cases} \alpha'_{ci} = \alpha_{ci} + N^{i}_{j,(.)} \\ \alpha'_{c} = \alpha_{c} + \sum_{i=1}^{K} N^{i}_{j,(.)} \\ \beta'_{c} = \beta_{c} + N^{K+1}_{j,(.)} \end{cases} \tag{94}$$

The topic-word update is defined as:

$$\begin{cases} \lambda'_{r} = \lambda_{r} + N^{i}_{(.),r} \\ \lambda' = \lambda + \sum_{r=1}^{V} N^{i}_{(.),r} \\ \eta' = \eta + N^{i}_{(.),V+1} \end{cases} \tag{95}$$

We can easily observe that the update equations obtained above from Eqs. 94 and 95 look very similar to those obtained in VB in the joint space of latent variables and model parameters (Eqs. 84 to 89). The multinomial updates are represented by $N^{i}_{j,(.)}$ for the document-topic count, and $N^{i}_{(.),r}$ for the topic-word count. In Eq. 96, we obtained the sampling equation of a topic z^{ij} in a particular document j in a class c given the observations $\mathcal{X}$ and the initial topic assignments associated to each word except the one being sampled z^{-ij}. From the collapsed Gibbs sampler, the multinomial variable $\hat{\psi}_{ijk}$ controls the counts in the document-topic and topic-word structures as in VB. However, the count $\hat{\psi}_{ijkc}$ in Eq. 96 obtained in the collapsed space is different from the one $(\tilde{\psi}_{ijk})$ in the joint space of VB. The collapsed representation offers in a particular class, the following update:

$$\hat{\psi}_{ijkc} = p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta) \tag{96}$$

using:

$$p(z_{ij}|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) = \frac{p(z_{ij}, z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)}{p(z^{-ij}, \mathcal{X}, c|\varepsilon, \zeta)} \tag{97}$$

from Eq. 90 so that:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) \propto \left(\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\lambda + \sum_{r=1}^{V} N^{-ij}_{.kr_{ij}})}{(\lambda + \eta + \sum_{r=1}^{V+1} N^{-ij}_{.kr_{ij}})} \right) \times \left(\frac{(\lambda_v + N^{-ij}_{.kv_{ij}})(\eta + N^{-ij}_{.k(V+1)_{ij}})}{(\sum_{r=1}^{V} N^{-ij}_{.kr_{ij}} + \lambda_r)} \right) \tag{98}$$

Now, the collapsed Gibbs sampler implements the Callen equations (Eq. 92) as in [12] to sample z given the observable variables $\mathcal{X}$. Consequently, in the collapsed space, the expected multinomial parameter in each class is estimated as a count from the true posterior distribution in the Eq. 96 while the VB updates its variational parameters in the joint space of the latent variables and model parameters using the expected multinomial parameter $\tilde{\psi}_{ijkc}$. This justifies again the difference between the two spaces as:

$$\tilde{\psi}_{ijkc} \neq \hat{\psi}_{ijkc} \tag{99}$$

3.3.3.3 The Collapsed variational Bayes (CVB) with the BL prior and Mean field variational inference

This new collapsed variational Bayesian inference (of the batch CVB-LBLA model) is naturally a combination of the BL-based VB and BL-based CGS. Therefore, it is a direct extension to the LDA and the CVB-LDA models as both still utilize the Dirichlet prior. Following the framework proposed in [12], the new BL-based CVB algorithm relaxes the strong independency assumption and provides a weaker assumption which is more accurate. It now models the dependence of parameters related to the latent variables in an exact fashion where parameters are either marginalized out or modeled separately as the joint $p(\theta, \varphi | z, \mathcal{X}, c, \varepsilon, \zeta)$. In both cases, it turns out it leaves the latent variables weakly dependent; as a result, assumed independent. Consequently, with this weak assumption, the BL-based CVB provides an efficient setting for mean field approximation as latent variables become conditionally independent given the parameters. With this conditionally independence assumption of the latent variables in the collapsed representation, a much better set of variational distributions could be obtained since the weaker assumption allows to decouple effectively the joint $\hat{\Delta}(z, \theta, \phi)$ as:

$$\hat{\Delta}(z, \theta, \varphi) = \hat{\Delta}(\theta, \varphi | z) \prod_{ij} \hat{\Delta}(z_{ij} | \hat{\psi}_{ij}) \tag{100}$$

where $\hat{\Delta}(z_{ij} | \hat{\psi}_{ij})$ is the variational multinomial distribution with parameters $\hat{\psi}_{ij}$ in the collapsed space, and the variational free energy $\hat{\mathcal{F}}(\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z))$ conditional to z becomes:

$$\hat{\mathcal{F}}(\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z)) = \mathbb{E}_{\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z)}[-\log p(\mathcal{X}, z, \theta, \varphi, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z)) \tag{101}$$

$$\hat{\mathcal{F}}(\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z)) = \mathbb{E}_{\hat{\Delta}(z)}[\mathbb{E}_{\hat{\Delta}(\theta, \varphi | z)}[-\log p(\mathcal{X}, z, \theta, \varphi, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{\Delta}(\theta, \varphi | z))] - \mathcal{H}(\hat{\Delta}(z)) \tag{102}$$

With only two variational posterior distributions ($\hat{\Delta}(\theta, \varphi | z)$, and $\hat{\Delta}(z)$), the variational free energy is minimized with respect to $\hat{\Delta}(\theta, \varphi | z)$, and then with respect to the collapsed variational $\hat{\Delta}(z)$ following the work in [12]. A minimum variational free energy is reached at the true posterior $\hat{\Delta}(\theta, \varphi | z) = p(\theta, \varphi | z, \mathcal{X}, c, \varepsilon, \zeta)$ which becomes :

$$\hat{\mathcal{F}}(\hat{\Delta}(z)) \triangleq \min_{\hat{\Delta}(\theta, \varphi | z)} \hat{\mathcal{F}}(\hat{\Delta}(z)\hat{\Delta}(\theta, \varphi | z)) = \mathbb{E}_{\hat{\Delta}(z)}[-\log p(\mathcal{X}, z, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{\Delta}(z)) \tag{103}$$

So, the bound in BL-based CVB in the batch CVB-LBLA can be expressed as :

$$-\log p(\mathcal{X} | c, \varepsilon, \zeta) \leq \hat{\mathcal{F}}(\hat{\Delta}(z)) = \mathbb{E}_{\hat{\Delta}(z)}[-\log p(\mathcal{X}, z, c | \varepsilon, \zeta)] - \mathcal{H}(\hat{\Delta}(z)) \tag{104}$$

$$\hat{\mathcal{F}}(\hat{\Delta}(z)) \leq \tilde{\mathcal{F}}(\tilde{\Delta}(z)) \triangleq \min_{\tilde{\Delta}(\theta)\tilde{\Delta}(\varphi)} \tilde{\mathcal{F}}(\tilde{\Delta}(z)\tilde{\Delta}(\theta)\tilde{\Delta}(\varphi)) \tag{105}$$

From Eq. 76, the optimization scheme using KL divergence in the collapsed space where the parameters θ and φ are marginalized out could therefore be reduced to:

$$(\hat{\psi}^*_{ij}) = \text{argmin}\, \hat{\psi}_{ij} D(\hat{\Delta}(z|\hat{\psi}_{ij})||p(z|\mathcal{X}, c, \varepsilon, \zeta, \mu)) \tag{106}$$

The Eq. 105 illustrates the BL-based CVB is a better and much improved approximation than the standard VB. This advantage is only provided in the collapsed representation [13]. In addition, minimizing the variational free energy $\hat{\mathcal{F}}(\hat{\Delta}(z))$ in Eq. 104 with respect to ψ_{ijk} leads to the multinomial update in each class as shown in Eq. 107 following the work in [12].

$$\hat{\psi}_{ijkc} = \hat{\Delta}(z_{ij} = k|c) = \frac{\exp(\mathbb{E}_{\hat{\Delta}(z^{-ij})}[\log p(\mathcal{X}, z^{-ij}, z_{ij} = k, c|\varepsilon, \zeta])}{\sum_{k'=1}^{K} \exp(\mathbb{E}_{\hat{\Delta}(z^{-ij})}[\log p(\mathcal{X}, z^{-ij}, z_{ij} = k', c|\varepsilon, \zeta])} \tag{107}$$

In CVB-LBLA model, the latent variables are drawn from the variational posterior distribution $\hat{\Delta}(z)$ using the BL-based CGS while the expected topic assignments lead to the parameters estimation when the Markov chain is stationary. The same conclusions are also found in [12, 67].

Using batch learning, the CVB-LBLA, for extremely large datasets, can be slow despite its accuracy in inference. In the literature, a solution to this handicap has been the use of Gaussian approximation or the second order Taylor approximation [12, 4]. The inference computes an extremely large amount of expectations. The batch learning's update in CVB-LBLA is finally computed as:

$$\begin{aligned}
\hat{\Delta}(z_{ij} = k|c) = \hat{\psi}_{ijkc} :\propto \\
&\left(\alpha_{ck} + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{jk.}]\right)\left(\lambda + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.k.}]\right) \\
&\times \left(\lambda_v + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.kx_{ij}}]\right)\left(\eta + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.k(V+1)_{ij}}]\right) \\
&\times \left(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.kr_{ij}}]\right)^{-1} (\mathbb{E}_{\hat{\Delta}}[N^{-ij}_{.k.}] + \sum_{r=1}^{V} \mathbb{E}_{\hat{\Delta}}[\lambda_r])^{-1} \\
&\times \exp\left(-\frac{Var_{\hat{\Delta}}(N^{ij}_{jk.})}{2(\alpha_k + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{jk.}])^2}\right) \\
&\times \exp\left(-\frac{Var_{\hat{\Delta}}(N^{ij}_{.k.})}{2(\lambda + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.k.}])^2}\right) \\
&\times \exp\left(-\frac{Var_{\hat{\Delta}}(N^{ij}_{.kx_{ij}})}{2(\lambda_v + \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.kxij}])^2}\right) \\
&\times \exp\left(+\frac{Var_{\hat{\Delta}}(\sum_{r=1}^{V+1} N^{ij}_{.kr_{ij}})}{2(\eta + \lambda + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{\Delta}}[N^{ij}_{.kr_{ij}}])^2}\right) \\
&\times \exp\left(+\frac{Var_{\hat{\Delta}}(N^{ij}_{.k.})}{2(\mathbb{E}_{\hat{\Delta}}[N^{ij}_{.k.}] + \sum_{r=1}^{V} \mathbb{E}_{\hat{\Delta}}[\lambda_r])^2}\right)
\end{aligned} \tag{108}$$

The sampling equation above shows that the CVB-LBLA samples its latent variables from a variational posterior distribution $\hat{\Delta}$ in the collapsed space of latent variables such that:

$$\hat{\psi}_{ij} = \sum_{k} \hat{\psi}_{ijk} \tag{109}$$

This observation is consistent with the work in [12], and the new update is also an improved version of the one estimated in that framework. The batch CVB-LBLA process is fully described in Algorithm 4.

3.3.3.4 Predictive distributions

At the stationary distribution, the batch CVB-LBLA's generative process for an unseen document utilizes its predictive distribution expressed in terms of its parameters conditional on the model hyperparameters. Using [12], the batch-based document parameter's distribution is estimated as:

$$\hat{\theta}_{jk} = \frac{(\alpha_{ck} + \mathbb{E}_{\hat{\Delta}}[N_{jk.}])}{(\mathbb{E}_{\hat{\Delta}}[N_{j..}] + \sum_{i=1}^{K} \alpha_{ck})} \tag{110}$$

Then conditional on the topic k, the predictive distribution of the words φ_{kw} is:

$$\hat{\varphi}_{kw} = \left(\frac{(\lambda + \mathbb{E}_{\hat{\Delta}}[N_{.k.}])(\lambda_v + \mathbb{E}_{\hat{\Delta}}[N_{.kx_{ij}}])}{(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{\Delta}}[N_{.kr_{ij}}])} \right) \times \left(\frac{(\eta + \mathbb{E}_{\hat{\Delta}}[N_{.k(V+1)_{ij}}])}{(\mathbb{E}_{\hat{\Delta}}[N_{.k.}] + \sum_{r=1}^{V} \lambda_r)} \right) \tag{111}$$

3.3.4 Evaluation method for the batch topic model

The topic model is always obtained using unsupervised learning. However, the lack of reliable topic labels for the dictionary codewords complicates ideas of using topics in a classification framework. It results in a need for an evaluation method that could assess or validate the robustness of the estimated topic model [18]. Since the goal is to compute efficiently the probability of the held-out dataset [42, 18], after estimation of the predictive distributions (parameters), we implemented the empirical likelihood estimate scheme presented in [18] as a validation method for the topics. In the batch CVB-LBLA model, the likelihood [12, 18] could be computed as:

$$p(\mathcal{X}) = p(\mathcal{X}|c, \varepsilon, \zeta) = \prod_{ij} \sum_{k} \hat{\theta}_{jk} \hat{\varphi}_{kw} \tag{112}$$

such that the counts $E_{\Delta}[N_{jk.}]$, $\mathbb{E}_{\Delta}[N_{.kv_{ij}}]$, and $\mathbb{E}_{\Delta}[N_{.kd_{ij}}]$ of the unseen document are obtained from the BL-based CVB sampling process in the collapsed space. As seen in Eq. 112, the parameters of the unseen document are then used to predict its likelihood [18]. However, the predictive likelihood $p(\mathcal{X}|c, \varepsilon, \zeta)$ is evaluated as follows: for an unseen document to be classified, some pseudo documents are generated with parameters θ using the BL priors from the training set. When the best candidates in documents in each class are obtained, we estimate their word probability distribution given the corpus parameter φ that leads to the class conditional probability $p(\mathcal{X}|c, \varepsilon, \zeta)$ [18]. With the class likelihood function, we can assess the probability of seeing the test set (unknown document) in the class. The class label is then given to the unseen document if it has the highest likelihood using the work presented in [2]. The empirical likelihood estimate is said to be robust compared to topic model's perplexity scheme as an evaluation method (validation) of the topic model [18].

The batch framework will be used when implementing the online learning that is based on accessing one minibatch at a time.

3.3.5 Classification's Bayesian decision boundary

The empirical likelihood estimate generates the probability of seeing the unseen document. In other words, it is used to get the class of the test set where the probability of seeing the class is proportional to the likelihood for a uniform class prior. Consequently, given an unseen document with its BoW representation $\mathcal{X}$, the probability of its class label (predictive model) is expressed following the Bayes rule as:

$$p(c|\mathcal{X}, \mu, \varepsilon, \zeta) \propto p(\mathcal{X}|c, \varepsilon, \zeta)p(c|\mu) \propto p(\mathcal{X}|c, \varepsilon, \zeta) \tag{113}$$

The decision about the category is ultimately made by the category label with the highest likelihood probability [2] such that:

$$C^* = \underset{c}{\operatorname{argmax}}\, p(\mathcal{X}|c, \varepsilon, \zeta) \tag{114}$$

3.3.6 Stochastic learning using minibatches: SCVB-LBLA and SCVB0-LBLA

Online learning is dominating Artificial Intelligence because in many situations the method is implemented when it becomes compuationally impossible to train over the entire dataset. In addition, many traditional models are still not capable of scaling to extremely large data. Online learning is also used to allow algorithms to dynamically adapt to new patterns in the data or in case of data that are time dependent. Nevertheless, it has been observed that a true online learning where the model updates the best predictor for future data at each step (by using one sample at the time) is not always possible as the model learns new inputs based on the current predictor value and the previous data. As the situation often requires to store all these previous data points, it is getting difficult to perform this technique. As a result, the constant space requirement is no longer guaranteed even though the time requirement to execute an update is fast.

A solution to this case has been the use of minibatches to maintain the constant memory requirement as the model is learning on a small batch instead. Though, in the past, a stochastic variational inference algorithm [65, 1] was implemented to scale the LDA inference to very large datasets. The scheme works on graphical models that operate with both global and local parameters for each data point x_j, and complete conditional distributions that are exponential family for each data variable [13]. In their online framework, the algorithm analyzes one data point at a time to learn about its local variational parameters such as θ_j that is used to update the global variational parameters such as the topics φ_k through a stochastic natural gradient update. As this method is guaranted to converge to the optimal variational solution [13], an extension of this approach to the batch CVB-LBLA is to update the parameter for the variational BL distribution on φ_k as shown in Eqs. 115 and 116 below. Using Eq. 75, we define the variational set $\tilde{\zeta}_k$ ($\tilde{\zeta}_k = (\tilde{\lambda}_{1k}, ..., \tilde{\lambda}_{Vk}, \tilde{\lambda}_k, \tilde{\eta}_k)$) as the parameter for the variational BL distribution on topic φ_k. Following [13], the updates on the parameter $\tilde{\zeta}_k$ involves a gradient update given as:

$$\tilde{\zeta}_k := (1 - \tau_t)\tilde{\zeta}_k + \tau_t \hat{\tilde{\zeta}}_k \tag{115}$$

This equation can therefore be extended as:

$$\begin{cases} \tilde{\lambda}_{vk} := (1 - \tau_t)\tilde{\lambda}_{vk} + \tau_t \hat{\tilde{\lambda}}_{vk} \quad for \;\; v = 1, ..., V \\ \tilde{\lambda}_k := (1 - \tau_t)\tilde{\lambda}_k + \tau_t \hat{\tilde{\lambda}}_k \\ \tilde{\eta}_k := (1 - \tau_t)\tilde{\eta}_k + \tau_t \hat{\tilde{\eta}}_k \end{cases} \tag{116}$$

Basically, by observing the generative process defined in section 4.2, the online scheme seems to compute in each document j, the variational distributions for the topic assignments and for document distributions over the topics using the VB method. The two equations above are similar to the oline EM algorithm proposed by [66] that maximizes the lower bound with respect to the parameter θ in M-step while the E-step provides a stochastic expectation step that updates the exponential family sufficient statistics using the online average framework as shown below:

$$\wp := (1 - \tau_t)\wp + \tau_t \hat{\wp}(\Omega_{n+1}; \theta) \tag{117}$$

such that Ω_{n+1} is a new data point, θ being the current parameter, and $\hat{\wp}(\Omega_{n+1}; \theta)$ is the estimate of the sufficient statistics based on the values of Ω_{n+1} and θ. Authors in [1, 65, 66] provided a platform to perform online variational inference in the collapsed space due to the flexibilities in that space such as computational speed, accuracy in inferences, and efficiency (convergence). Therefore, this section will deal with the stochastic variational inference using the collapsed representation already presented in the batch CVB-LBLA in the previous sections.

As the amount of data being processed in online fashion can be extremely large and close to infinity, it is clear that the fundamental goal in streaming is to avoid storing all previous data in contrast to the batch method. As a result, iteratively, online setting only stores current data or mini-batch and their associated topic assignments while the rest of the data are neglected. The time complexity of the CGS in one iteration is in the order of $O(K|\Upsilon|)$ where $|\Upsilon|$ is the size of the mini-batch following the framework in [13]. The way the algorithm is able to slowly prune out the previous data (as if it was forgetting the history in the data) is through the use of the decay factor τ_t as seen from Eqs. 115 to 117. For instance, as the CGS stores the sufficient statistic $\hat{N}_{kv}$ which emphasizes on the topic-word counts, the optimized online learning framework insures that the count $\tau_t \hat{N}_{kv}$ helps the model forgetting the history, such that the posterior distribution of the previous data becomes weaker in every iteration.

One of the problems in the batch is that every token is associated with a variational distribution ψ_{ij} which in overall affects memory space [13]. In online setting though, the scheme is designed in a way that does not allow to substract current value of ψ_{ij} as usually observed in Eq. 108 with the collapsed Gibbs sampler. This is because the true online algorithm does not store the values of ψ, but only estimates their update versions. Most importantly, for large scale processing, removing just a current value of ψ_{ij} compared to the remaining in the corpus does not have any significant impact on the overall update equation (Eq. 108) as $N^{-ij} \approxeq N^{ij}$. Though, the collapsed variational Bayesian inference (CVB) does maintain variational distributions ψ_{ij} over its K topics for each word i and document j. As the optimization of the lower bound with respect to the variational distribution ψ_{ij} is not tractable, it has been shown a scheme in [12] where approximating the updates does work better in practice. It does outperform the VB in prediction. Then, authors in [4] presented the simpler version (CVB0) of the CVB performing faster than the CVB itself while still maintaining the accuracy of the CVB. The CVB0 is shown in Eq. 108 and according to [4], it is the fastest technique for LDA inference for single core batch inference in terms of convergence rate. The CVB algorithm has update that is deterministic [40] since it carries all the advantages of the VB. It shows the algorithm iteratively updating each ψ_{ij} where CVB0 statistics are estimated as: $N_Z^k \triangleq \sum_{ij} \psi_{ijk}$; $N_\theta^{jk} \triangleq \sum_i \psi_{ijk}$; and $N_\varphi^{wk} \triangleq \sum_{ij:w_{ij}=w} \psi_{ijk}$. Though, a disadvantage of the CVB0 is the extremely large memory required to store the variational distribution over each token in the corpus. This leads to the stochastic CVB0 or SCVB0 as a solution for memory space problem. This inference can be obtained by constructing the stochastic SCVB-LBLA, our online model.

3.3.6.1 The BL-based SCVB0

In our model, we are implementing an online or stochastic collapsed variational Bayesian inference for the latent Beta-Liouville allocation (SCVB-LBLA) model following the online topic modeling framework that has been developped in [1, 66, 13]. As presented in the previous sections, these authors laid out the foundations for online learning using the collapsed represenation in the original LDA model. As a result, we are proposing an extension to their work using the new LBLA model in the collapsed space since we are highly motivated by the flexibilities of the Beta-Liouville prior.

Following the stochastic method in [13], for a uniform and random draw of a token in the corpus, using the variational sampling distribution Δ, some expectations of the sufficient statistics could be estimated such as: $\mathbb{E}_\Delta[N_{.k.}] = \Lambda\psi_{ij}$ where Λ is the number of words in the corpus while $\mathbb{E}_\Delta[N_{.kx_{ij}}] = \Lambda\Psi^{(ij)}$ where $\Psi^{(ij)}$ is the $V \times K$ matrix, and $\mathbb{E}_\Delta[N_{j..}] = \Lambda_j\psi_{ij}$. Using the compact representation (Eq. 118) of the sufficient statistics in [13], we can really see and appreciate the contribution of the CVB-LBLA and the SCVB-LBLA by observing Eqs. 108 and 118. It shows the extra terms in the updates that are document and topic specific to the SCVB-LBLA and its batch CVB-LBLA only. This makes the difference between the LDA and our proposed model.

$$\begin{cases} \mathbb{E}_\Delta[N_{j..}] = \mathbb{E}_\Delta[N_\theta^j] \\ \mathbb{E}_\Delta[N_{k_{x_{ij}}}] = \mathbb{E}_\Delta[N_\varphi] \\ \mathbb{E}_\Delta[N_{.k.}] = \mathbb{E}_\Delta[N_Z] \end{cases} \tag{118}$$

For convenience, the exponential term in the update equation (Eq. 108) has been neglected without affecting the estimates as suggested in [13]. Importantly, the topic and word parameters are still drawn for the BL. As the variational distributions ψ cannot be maintained, it is difficult to directly perform the sampling process. However, using a current guess of the CVB0 statistics, an update of a word variational distribution could be provided to allow observation of its new value. In this iterative procedure, as the values of ψ_{ij} changes everytime, the traditional simple average is no longer possible. The scheme ultimately uses the online average of the statistics following the work proposed in [66] as illustrated in Eq. 117. Therefore, since ψ are not stored, and in practice it is too expensive to update the entire expected N_φ (sufficient statistics) for every token, a solution is the use of minibatches and minibatch updates in order to maintain the constant memory space requirement for the online technique. The expected $N_{k_{x_{ij}}}$ after observing a mnibatch Υ is the average of the per-token estimates. This leads to the following update in $N_{k_{x_{ij}}}$ and $N_{.k.}$ using Eqs. 119 to 123:

Algorithm 4 summary of the batch based CVB-LBLA Inference

```
procedure
Input: $\mathcal{X}$, $(\varepsilon, c) = (\alpha_{c1}, ..., \alpha_{cK}, \alpha_c, \beta_c)$, $iterMax$, $\zeta = (\lambda_1, ..., \lambda_V, \lambda, \eta)$, $K$, $V$, $N$
Initialize $z$, $N_{jk.}$, $N_{.kx_{ij}}$, $N_{.k.}$
    for $iter = 1$ to $iterMax$ do
        for $i = 1$ to $N$ in document $j$ in class $c$ do
            $z_{ij} \sim \Delta(z_{ij} = k | z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta)$ using Eq.108
            Update the counts $N_{.k.}$, $N_{jk.}$, and $N_{.kx_{ij}}$
        end for
    end for
Output: Parameters $\tilde{\theta}_{jks}$ and $\tilde{\varphi}_{kws}$ using Eq.147 and 148
end procedure
```

In the local update (document parameters), we have:

$$N_{j..} := (1 - \tau_{t_\theta}) N_{j..} + \tau_{t_\theta} \hat{N}_{j..} \tag{119}$$

But in a uniform draw of a token, the equation becomes:

$$N_{j..} := (1 - \tau_{t_\theta}) N_{j..} + \tau_{t_\theta} \Lambda_j \hat{\psi}_{ij} \tag{120}$$

Concerning the global update (corpus parameters), we obtain:

$$N_{k_{x_{ij}}} := (1 - \tau_{t_\varphi}) N_{k_{x_{ij}}} + \tau_{t_\varphi} \hat{N}_{k_{x_{ij}}} \tag{121}$$

$$N_{k(V+1)} := (1 - \tau_{t_\varphi}) N_{k(V+1)} + \tau_{t_\varphi} \hat{N}_{k(V+1)} \tag{122}$$

$$N_{.k.} := (1 - \tau_{t_\varphi}) N_{.k.} + \tau_{t_\varphi} \hat{N}_{.k.} \tag{123}$$

where $\hat{N}_{k_{x_{ij}}} = \frac{\Lambda}{|\Upsilon_m|} \sum_{ij \in \Upsilon} \Psi^{(ij)}$, $\hat{N}_{.k.} = \frac{\Lambda}{|\Upsilon_m|} \sum_{ij \in \Upsilon} \psi_{ij}$, and $\hat{N}_{k(V+1)} = \frac{\Lambda}{|\Upsilon_m|} \sum_{ij \in \Upsilon} \Psi^{(V+1)}$ such that $|\Upsilon_m|$ is the size of the mth minibatch in the database.

The SCVB-LBLA's online inference scheme that takes advantage of the online average framework is summarized in Algorithm 5.

3.4 Experimental results

In the following experiments, we have implemented several challenging applications using the stochastic model SCVB-LBLA. We mainly compared our new approach to its batch-based competitors and some other online schemes such as the SCVB-LDA. These applications include text analysis, images and videos recognition and categorization. In this framework, the testing data are set in online fashion where the samples (minibatches in this case) arrive sequentially, one at a time. This is in contrast to the batch learning methods (CVB-LBLA and CVB-LDA) that process all the available data at the same time.

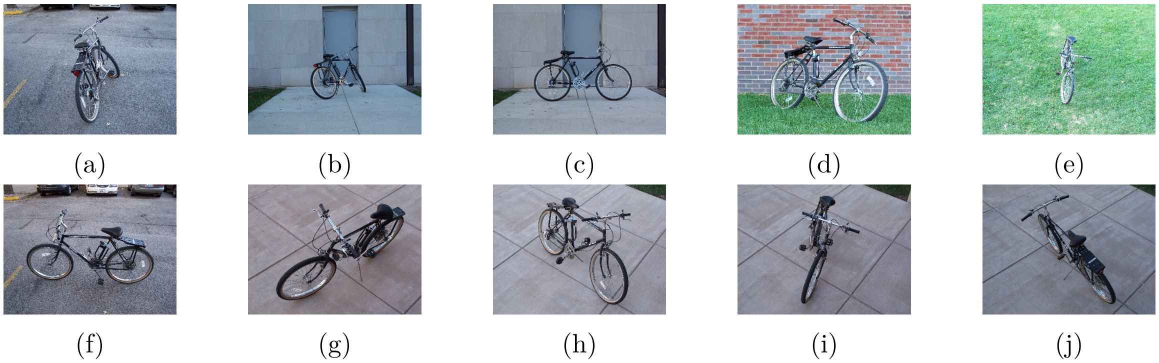
Following the bag of visual words scheme, local features provide the representation of both the training and the testing sets using the count data (frequency counts). This representation is needed for models such as the SCVB-LBLA since they are usually effective with count data. To speed up computations, the exhaustive search for the optimal number of topics, and vocabulary size following the work in [67] is a little relaxed by fixing the size of the vocabulary while varying only the topics.

Algorithm 5 summary of the stochastic SCVB-LBLA Inference

```
procedure
•Initializations: $N_{.kx_{ij}}$, $N_{j..}$, $N_{.k.} := \sum_w N^w_{.kx_{ij}}$, $\tau_{t_\theta}$, $\tau_{t_\varphi}$, $iterMax$, $burnInIter$

    for $iter$ = 1 to $iterMax$ do
        for $m$ = 1 to $\Upsilon_M$ (for each Minibatch $\Upsilon_m$) do
         ◇ $\hat{N}_{.kx_{ij}} := 0$ , $\hat{N}_{.k.} := 0$

            for $j$ = 1 to $|\Upsilon_m|$ (for each document $j$ in $\Upsilon_m$) do

                if $iter < burnInIter$ ("Burn − in"process) then
                    • Update $\hat{\psi}_{ij}$ using Eq.108
                    • Update $N_{j..}$ using Eq.119

                end if
                if $iter \geq burnInIter$ (for each token $i$) then
                    ◇ Update $\hat{\psi}_{ij}$ using Eq.108
                    ◇ Update $N_{j..}$ using Eq.119
                    ◇ Compute $\hat{N}^{x_{ij}}_{.kx_{ij}} := \hat{N}^{x_{ij}}_{\varphi} + \frac{\Lambda}{|\Upsilon_m|}\hat{\psi}_{ij}$
                    ◇ Compute $\hat{N}^{x_{ij}}_{.k.} := \hat{N}^{x_{ij}}_{.k.} + \frac{\Lambda}{|\Upsilon_m|}\hat{\psi}_{ij}$
                    ◇ Compute $\hat{N}^{(V+1)}_{.kx_{ij}} := \hat{N}^{(V+1)}_{.kx_{ij}} + \frac{\Lambda}{|\Upsilon_m|}\hat{\psi}_{ij}$

                end if
        ◇ Update $N_{.kx_{ij}}$ in Eq.121
        ◇ Update $N_{.k.}$ in Eq.123
        ◇ Update $N^{(V+1)}_{.kx_{ij}}$ in Eq.122

            end for
        end for
    end for
end procedure
```

	Natural Scenes Images	Face expressions	Cohn Kanade data	Action Recognition in videos
CVB-LDA	64.13%	65.51%	65.12%	64.86%
SCVB-LDA	68.56%	63.41%	60.96%	62.65%
SCVB-LBLA	76.8%	74.29%	77.14%	72%
CVB-LBLA	70.34%	65.8%	63.61%	66.25%

Table 3.2: Comparison between the new SCVB-LBLA model and the other schemes within the BoW framework



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 3.1: An object from a bicycle's class at different 2D views for a 3D modeling

3.4.1 Online Image categorization for Natural scenes dataset

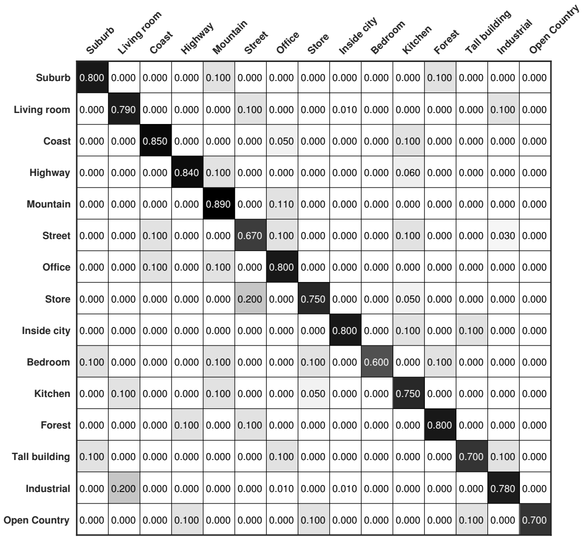
3.4.1.1 Methodology

In this experiment, we used the well-known grayscale natural scenes dataset [48] in online fashion to recursively update the class distribution as we grouped the corpus documents or images into a set of minibatches. Using Fig. 2.3 and Table 3.3, this challenging dataset has 15 categories that include suburb, living room, coast, forest, highway, mountain, street, office, store, bedroom, inside city, tall building, open country, kitchen, and industrial. In each category, the data is divided into a testing set that contains 100 samples while the remaining constitutes the training set. The minibatch contains around 10 images. In each class, the BoW representation of the corpus (Fig. 3.12) leads to a vector of counts in each document (image) after the codebook formation using the framework in [2] with the SIFT (Scale Invariant Feature Transform) feature [71]. The training set is used to implement the SCVB-LBLA model with asymmetric BL priors. Using the minibatch, the model's parameters estimation provides the predictive model. And with these predictive distributions, the empirical likelihood framework is constructed to evaluate the topics. It then leads to the class likelihood estimation which helps predicting the class label of unseen images or documents in the minibatch. As a result, the category of an unseen image is chosen by the class with the highest class posterior distribution that is equivalent to the class conditional probability for a uniform class prior in our case. For the online average, we used a step size or learning rate (τ_t) of 0.1 for the document-topic update and for the word-topic topic update.

The SCVB-LBLA showed an overall classification accuracy of 76.8% (Fig. 3.2) with a very short (fast) runtime of 7 min (minutes). The optimal number of topics is $K = 95$. These results could be explained by the advantages provided by the collapsed representation as now the model has ability to estimate with high efficiency the posterior distribution due to the stability of lower bound in the collapsed space. It has been shown in the literature [12, 13] that the collapsed space provides a better approximation than the uncollapsed space. These results also demonstrate the flexibility of the BL prior (general covariance structure as in Fig. 3.3) compared to the traditional Dirichlet distribution that is very limited for its inability to perform in case of positively correlated datasets.

Categories	Size
suburb	241
living room	289
cost	360
forest	328
highway	260
mountain	374
street	292
office	215
store	315
Bedroom	216
Inside City	308
Tall buidling	356
Open country	410
Kitchen	210
Industrial	311

Table 3.3: size of each image category.



	Suburb	Living room	Coast	Highway	Mountain	Street	Office	Store	Inside city	Bedroom	Kitchen	Forest	Tall building	Industrial	Open Country
Suburb	0.800	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000
Living room	0.000	0.790	0.000	0.000	0.000	0.100	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.100	0.000
Coast	0.000	0.000	0.850	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000
Highway	0.000	0.000	0.000	0.840	0.100	0.000	0.000	0.000	0.000	0.000	0.060	0.000	0.000	0.000	0.000
Mountain	0.000	0.000	0.000	0.000	0.890	0.000	0.110	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Street	0.000	0.000	0.100	0.000	0.000	0.670	0.100	0.000	0.000	0.000	0.100	0.000	0.000	0.030	0.000
Office	0.000	0.000	0.100	0.000	0.100	0.000	0.800	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Store	0.000	0.000	0.000	0.000	0.000	0.200	0.000	0.750	0.000	0.000	0.050	0.000	0.000	0.000	0.000
Inside city	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.800	0.000	0.100	0.000	0.100	0.000	0.000
Bedroom	0.100	0.000	0.000	0.000	0.100	0.000	0.000	0.100	0.000	0.600	0.000	0.100	0.000	0.000	0.000
Kitchen	0.000	0.100	0.000	0.000	0.100	0.000	0.000	0.050	0.000	0.000	0.750	0.000	0.000	0.000	0.000
Forest	0.000	0.000	0.000	0.100	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.800	0.000	0.000	0.000
Tall building	0.100	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.700	0.100	0.000
Industrial	0.000	0.200	0.000	0.000	0.000	0.000	0.010	0.000	0.010	0.000	0.000	0.000	0.000	0.780	0.000
Open Country	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.100	0.000	0.700

Figure 3.2: Confusion matrix for the natural scenes classification

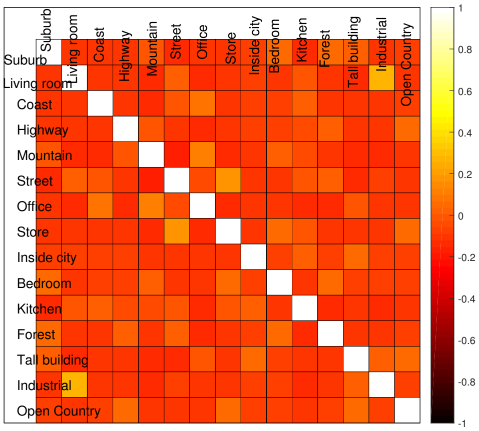


Figure 3.3: Natural scene images correlation map.

3.4.2 Online Facial Expression recognition and categorization

3.4.2.1 JAFFE dataset

Driven by social media, facial expressions modeling and sentiment analysis are hot topics today in the field of Artificial Intelligence [72, 49, 50]).

Concerning facial expressions, the modeling focuses on the intrinsic characteristics of the facial textures (Fig. 3.6). In this particular application, we used the JAFFE (Japanese Female Facial Expression) dataset (Figs. 3.4 and 3.5). It has 213 images collected from 10 Japanese females showing 7 facial expressions [73, 74, 75] such as anger, sadness, surprise, happiness, fear, disgust, and neutral. The first task is to group these females according to these seven expressions representing therefore our different classes. Following the method proposed in [72, 76], we preprocessed (cropping) each image to obtain 189 × 100. This is because each original image has size of 256 × 256. Therefore, this crop does allow to focus more on the facial characteristics as we discard a lot of the background image to guarantee better feature representation. So, we divided each image into 90 (9 × 10) blocks or regions where each region is 14 × 15 pixels. Then, these regions textures are represented using the 59-bin LBP operator in the $(8, 2)$ neighborhood as it provides 8 samples (neighbors) on a circle of radius 2) following the framework in [51, 59]). It leads to a histogram of length $(90 \times 59) = 5310$. As the vector is very long, to avoid overfitting problem, we used a pLSA model from the framework presented in [77] for dimensionality reduction at different sizes such as $K = 48, 64, 80$, and 128. We finally retained the most relevant 128 elements of proportions from the original 5310−dimensional vector. The training set being represented using the LBP feature vectors, we trained our SCVB-LBLA model where each minibatch contained around 5 images.

The confusion matrix in Fig. 3.7 shows a classification accuracy rate of 74.29% which outperforms its competitors (see Table 3.2). The minibatch online scheme is very fast with an optimal number of topics around $K = 52$. The model was able to perform the training and the testing tasks in less than 15 min. The CVB-LDA and CVB-LBLA were both slow compared to the SCVB-LBLA. This again shows the flexibility of the BL prior as it has a full covariance structure and less parameters than the GD that help in fast computation. In addition, the LBP descriptor was robust as it provided the relevant features for an accurate inference coupled with a stable lower bound that resulted in a better performance in overall.

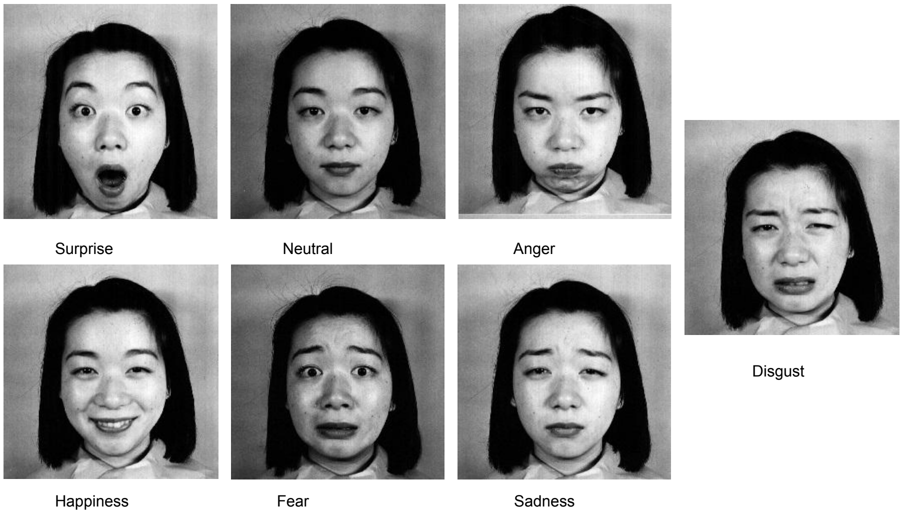


Figure 3.4: Facial expressions and emotions in the JAFFE dataset

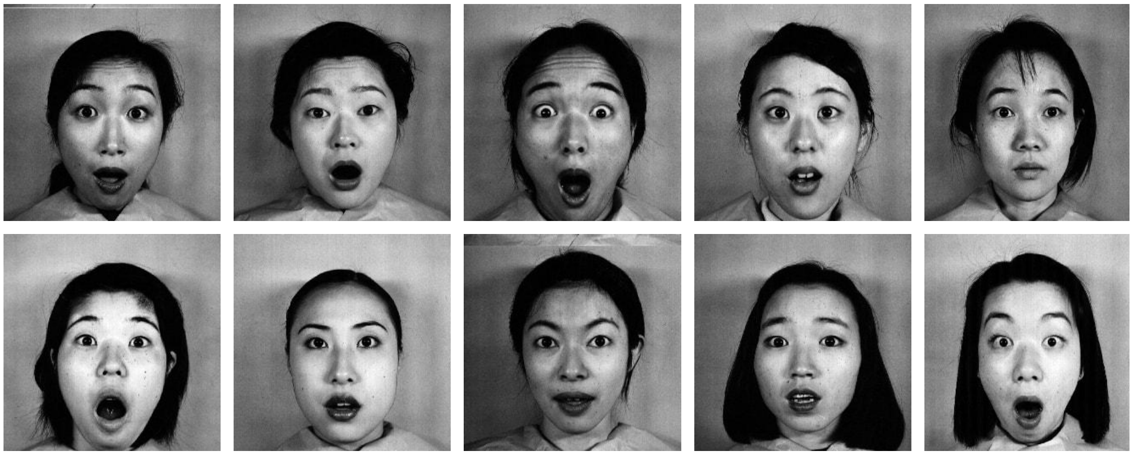


Figure 3.5: Women showing a "surprised" facial expression

3.4.2.2 Cohn-Kanade dataset

Cohn-Kanade database [78, 79] contains facial expressions of some males and females, 97 individuals in total. This is a collection of 486 images sequences of about 1.70 GB (Fig. 3.8). The resolution of each image is 640 × 490. So, in the preprocessing, following the same scheme for the JAFFE dataset, we cropped the original images again leaving only the relevant parts of facial characteristics. We obtained a small image of size 200 × 100 which led to a 50 blocks from a 20 × 20 pixels per block. Using the 59-bin LBP operator we got an histogram of size $59 \times 50 = 2950$ from which we finally obtained the reduced 128 dimensional vector from the pLSA algorithm [77].

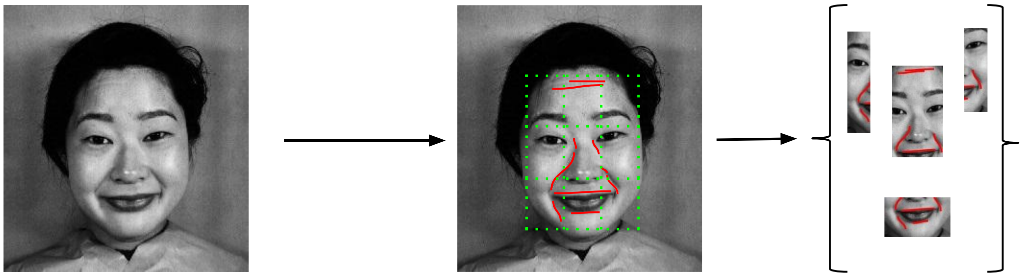
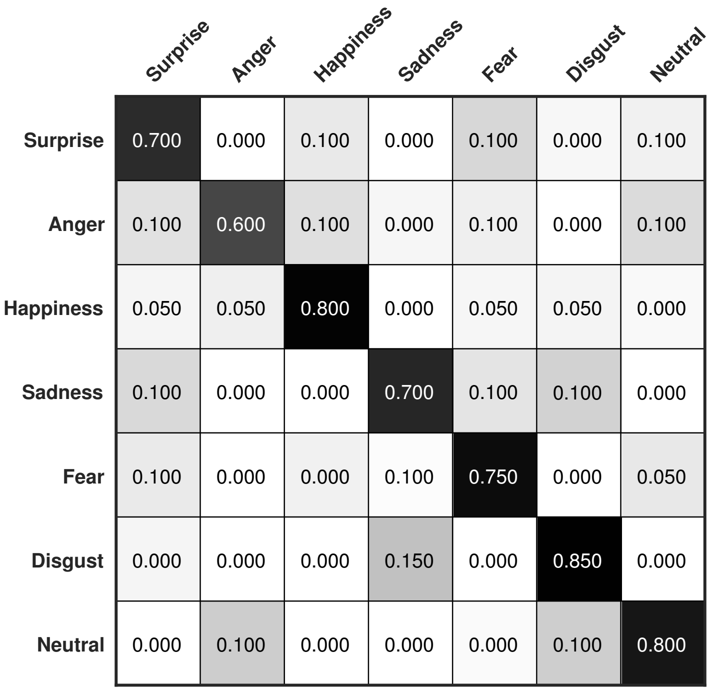


Figure 3.6: Facial Expression: Key Regions of Interest and Extraction



	Surprise	Anger	Happiness	Sadness	Fear	Disgust	Neutral
Surprise	0.700	0.000	0.100	0.000	0.100	0.000	0.100
Anger	0.100	0.600	0.100	0.000	0.100	0.000	0.100
Happiness	0.050	0.050	0.800	0.000	0.050	0.050	0.000
Sadness	0.100	0.000	0.000	0.700	0.100	0.100	0.000
Fear	0.100	0.000	0.000	0.100	0.750	0.000	0.050
Disgust	0.000	0.000	0.000	0.150	0.000	0.850	0.000
Neutral	0.000	0.100	0.000	0.000	0.000	0.100	0.800

Figure 3.7: Confusion matrix from the JAFFE Facial expressions classification

Similar to the JAFFE dataset, the Cohn-Kanade data is also classified using emotions such as Anger, Disgust, fear, Joy, Sadness, Surprise, and Neutral. The minibatches from the BoW allow to perform the online scheme. We obtained a classification accuracy of 77.14% (Fig. 3.9) with a time frame of 20 min. This is a very challenging dataset. The batch version of our model (CVB-LBLA) was very slow on this data including the CVB-LDA. It demonstrates the advantage of the online framework where the model does not require to access all the documents, but still can provide very accurate updates. Finally, the use of a flexible prior which can perform in full topic correlation framework has been beneficial to the model's performance (Fig. 3.10) that reaches an accuracy of 77.14% at $K = 90$ in Fig. 3.11.

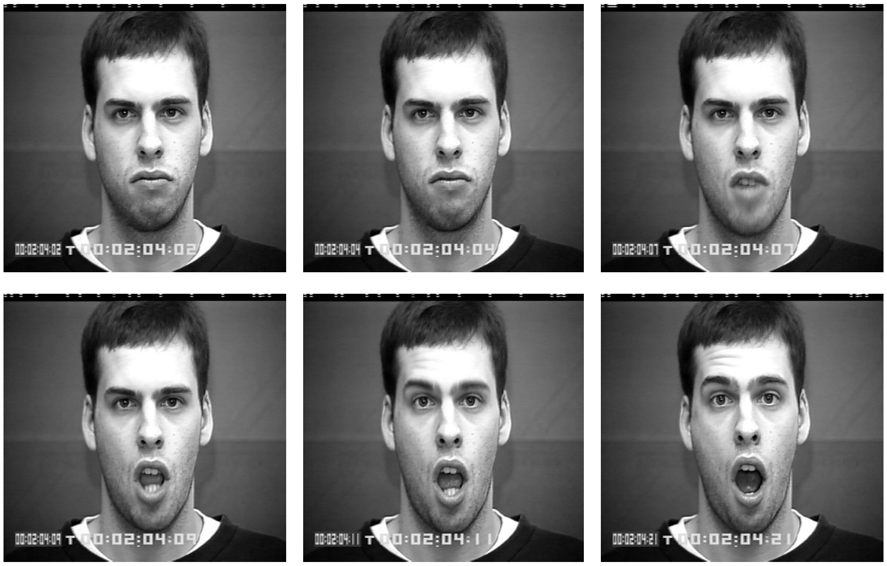


Figure 3.8: Facial emotions in the Cohn Kanade dataset

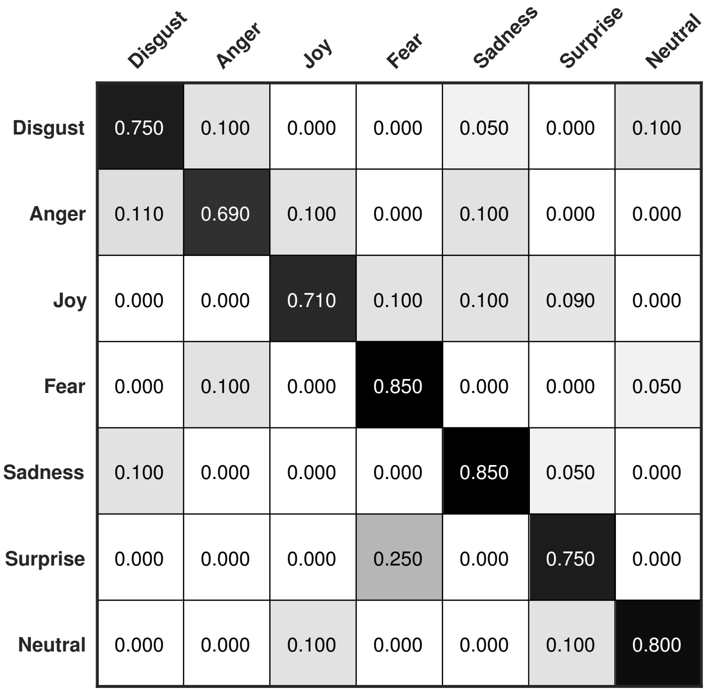
3.4.3 Online Action recognition in videos

In this experiment, we implemented the action recognition in videos using the KTH dataset that contains 2391 video sequences at 25 frames [54, 55]. We used the optical flow algorithm to collect relevant features for the BoW representation of the corpus data. The dataset is mainly comprised of 25 individuals in 4 scenarios performing 6 types of human actions such as running, walking, jogging, hand waving, boxing, and hand clapping as shown in Table 3.6. Concerning this table, each column represents a human action in these 4 different scenarios. For processing purpose, the sequences were downsampled into a resolution of 160 by 120 pixels with a length of 4 seconds. In our experiment, 60% of the dataset were used for training while the remaining constitutes the testing set. Approximately 100 frames were collected from each video sequence in each class with a minibatch size of 10 per class.

The BoW framework allows the transformation of the features collected from the optical flow scheme ([56]) into frequency counts following the method in [35]. These count data are then used to construct the stochastic SCVB-LBLA to learn the minibatches in online fashion. The model provided an overall classification accuracy of 72% (Fig. 3.13) with a runtime around 18 min; which suggested the SCVB-LBLA was robust and faster than the CVB-LBLA and CVB-LDA because when we implemented these two batch models using the same dataset, their runtime was considerably higher. The optimal number of topics $K = 150$. In general, this is a computationally expensive learning that could be impossible to obtain satisfying results without an online framework built with a flexible prior in a proper modeling space (collapsed space) that perform in a topic correlation (Fig. 3.18) environment for possible model selection as seen in Fig. 3.11 (for the Cohn-Kanade dataset).

3.4.4 Online text processing

In this application, three datasets have been used: First, the KOS (www.dailykos.com) which has in total $J = 3430$ documents for a vocabulary size of $W = 6909$. The total



	Disgust	Anger	Joy	Fear	Sadness	Surprise	Neutral
Disgust	0.750	0.100	0.000	0.000	0.050	0.000	0.100
Anger	0.110	0.690	0.100	0.000	0.100	0.000	0.000
Joy	0.000	0.000	0.710	0.100	0.100	0.090	0.000
Fear	0.000	0.100	0.000	0.850	0.000	0.000	0.050
Sadness	0.100	0.000	0.000	0.000	0.850	0.050	0.000
Surprise	0.000	0.000	0.000	0.250	0.000	0.750	0.000
Neutral	0.000	0.000	0.100	0.000	0.000	0.100	0.800

Figure 3.9: Confusion matrix from Cohn-Kanade Facial expressions classification

number of words in the corpus is $N = 467,714$. So on average, the dataset has 136 words per document. The second one is the NIPS (books.nips.cc) dataset with $J = 1675$ documents, a vocabulary size of size $W = 12419$, and a total words of $N = 2,166,029$ in the corpus. It has on average 1293 words per document. These two datasets are used following the work in [12]. Then finally, we have PubMed Central dataset which is a corpus of 320 millions from 165,000 articles (documents) with a vocabulary size of around 38,500 words. It is presented in [13].

Different from the previous applications that use images in a classification framework, now we are only interested in analyzing text documents using the online model to show its flexibility with extremely large datasets that could even challenge batch models for storage issues. In Table 3.4, it shows that the SCVB-LBLA outperforms its competitors in every dataset regardless of the size in terms of processing speed. The minibatches obtained here were larger due to the size of the corpus. However, the stochastic model is just faster than the other schemes. The online framework does not need or is not required to access all the data before providing estimates and updates. The batch in constrast does require the use of all the available data in order to compute updates. This is therefore translated into the significantly increase runtime in the batch compared to the online (decrease). The CVB-LBLA's performance is getting slower and slower as the data increase in size. In addition, in terms of convergence, the SCVB-LBLA and the SCVB-LDA outperform the batch models such as the CVB-LDA and the CVB-LBLA as shown from Figs. 3.14 to 3.17. Though, as we notice in these figures, the SCVB-LBLA is the fastest.

Fig. 3.13 shows the performance of the model in identifying and classifying actions in a video sequence. Fig. 3.14 illustrates the performance of the new approach in terms

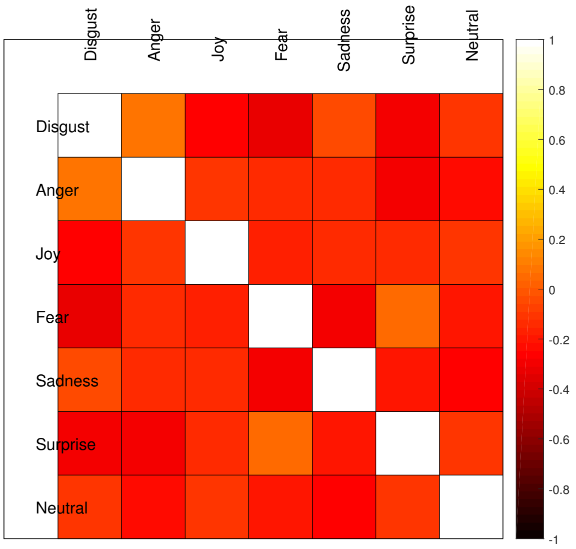


Figure 3.10: Cohn-Kanade dataset's correlation map

of time complexity and its easy access to convergence: as seen, the number of iterations to build the model is smaller in our proposed approach, which means that our method is faster than its competitors. In other words, it requires very few iterations to reach convergence in the new scheme. The batch-based techniques (CVB-LDA and CVB-LBLA) are slower to reach convergence than the stochastic approaches (SCVB-LDA and SCVB-LBLA). And the SCVB-LBLA remains the fastest technique. Similar observations and conclusions could be made from analyzing Figs. 3.15 to 3.17 using a variety of datasets such as texts (NIPS data), images (natural scenes), and videos (KTH datasets). It shows again that the stochastic approach outperforms the batch method. Still, the SCVB-LBLA is the dominant performer compared to its competitors. Fig. 3.18, through the correlation map, illustrates the dependency between distinct random variables (classes) in our document classification problem. It also shows the flexibilities of the BL prior in handling negatively and positively correlated datasets.

Finally, Table 3.5 provided the robustness of our new approach compared to some recent competitors under the same dataset (the natural scene categorization dataset). The table also illustrates the major characteristics of each model. Coupled with the BL priors, the proposed model with its flexibility and efficiency was able to handle all the 15 categories in the dataset. The batch method proposed in [2] performed on 13 categories while the online framework in [64] used 7 classes. Nevertheless our proposed approach has provided a better accuracy when considering the number of categories used by each method. In addition, the collapsed representation in the SCVB-LBLA has considerably helped in the estimation. Because the CVB is a combination of the VB (variational Bayes) and the CGS (collapsed Gibbs sampler), our hybrid yielded deterministically (convergence due to the VB) results (estimates) that are accurate (owing to the CGS) in online fashion.

Our online scheme through its performance could be seen as the the favored candidate compared to the batch-based techniques by allowing data to be processed one at a time. Therefore, the technique facilitates complex data handling while allowing a better storage

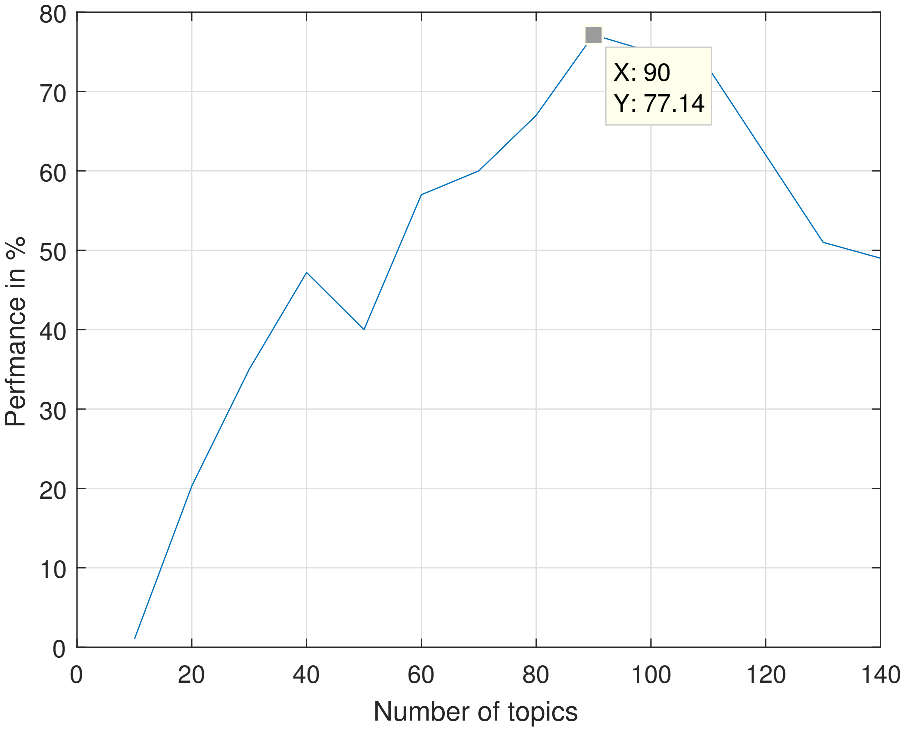


Figure 3.11: Optimal number of topics for the Cohn-Kanade dataset

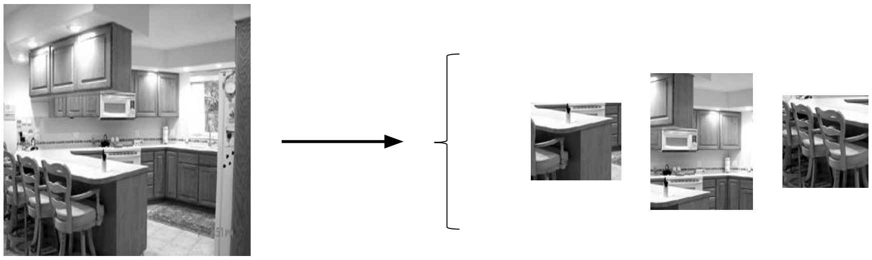
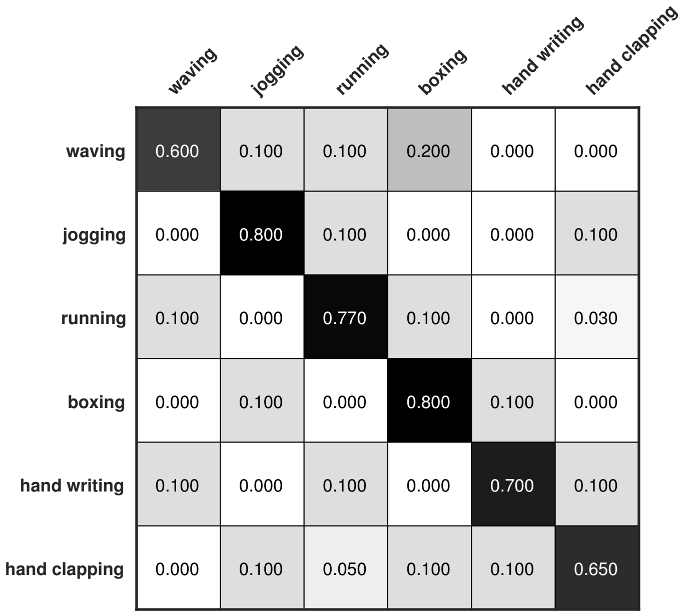


Figure 3.12: Natural scene image Features extraction

management and computational speed.

3.5 Conclusion

In online learning scheme, we implemented a technique that used the BL instead of the Dirichlet in the collapsed representation (that provided useful properties for stochastic methods). As the GD (generalized Dirichlet distribution), the BL is a generalization of the Dirichlet distribution. However, it has less parameters compared to the GD. As a result, the BL has been seen as a good candidate for models that provide fast computations. The



	waving	jogging	running	boxing	hand writing	hand clapping
waving	0.600	0.100	0.100	0.200	0.000	0.000
jogging	0.000	0.800	0.100	0.000	0.000	0.100
running	0.100	0.000	0.770	0.100	0.000	0.030
boxing	0.000	0.100	0.000	0.800	0.100	0.000
hand writing	0.100	0.000	0.100	0.000	0.700	0.100
hand clapping	0.000	0.100	0.050	0.100	0.100	0.650

Figure 3.13: Confusion matrix of the action classes in video

Text Data	CVB-LBLA	CVB-LDA	SCVB-LDA	SCVB-LBLA
NIPS	8 min	17 min	11 min	5 min
KOS	35 min	48 min	20 min	12 min
PubMed Central Times	95 min	120 min	75 min	42 min

Table 3.4: Batch CVB-LBLA and the stochastic SCVB-LBLA in terms of the runtime in minutes for different data sizes

stochastic framework we set has also provided through the performance of our new approach, stability, accuracy in inferences, and efficiency in convergence. The SCVB-LBLA has an improved time complexity due to the fact that it operates on minibatches in online fashion. The scheme is memoryless, and the decay factor (learning rate) truly helps mitigating the influence of the old data in favor of the new ones as it performs the online average technique on the sufficient statistics of the model. Consequently, the new technique is able to provide a solution to the memory space management's problem often observed in the batch-based models. All these advantages in the stochastic SCVB-LBLA are due to the flexibility of the conjugate prior that allows both topic correlation and vocabulary analysis in these datasets. The general covariance structure in the BL is also suitable for any data modeling within the BoW framework. This is not the case for the Dirichlet for its limitation in case of positively correlated datasets. Our method presented in this chapter outperforms the batch models including the Dirichlet-based online schemes. This ultimately shows the robustness of the new approach as it is flexible to so many data types including nonstationary datasets.

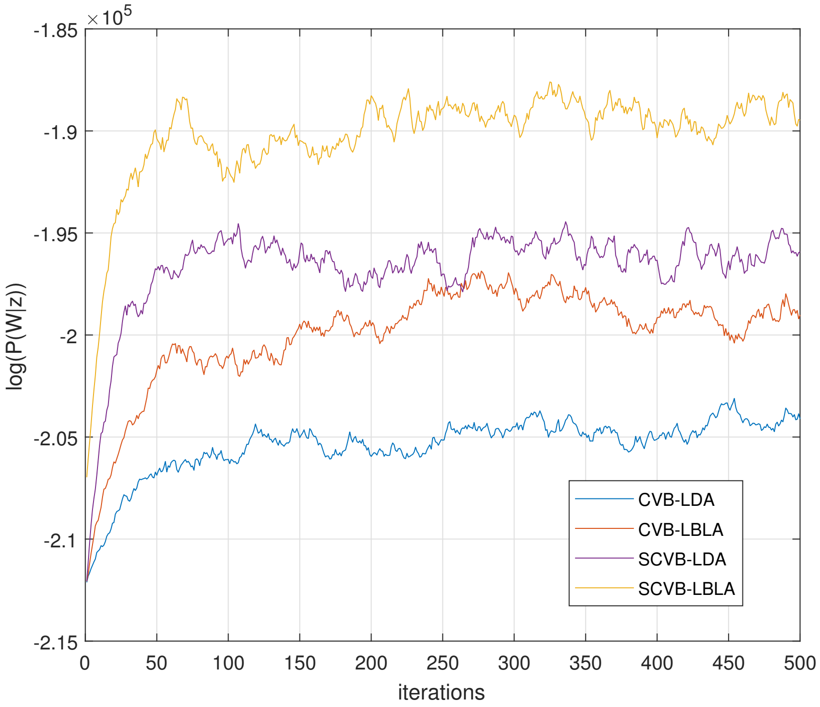


Figure 3.14: Convergence Process between the SCVB-LBLA and its competitors using text documents

Future work could still cover another extension to LDA that combines two different flexible priors (Beta Liouville and generalized Dirichlet distributions, for instance) in the generative process using the collapsed representation for both the batch and stochastic frameworks. We could also investigate on using other powerful feature extraction schemes that could enhance analysis for better detection, recognition, and classification.

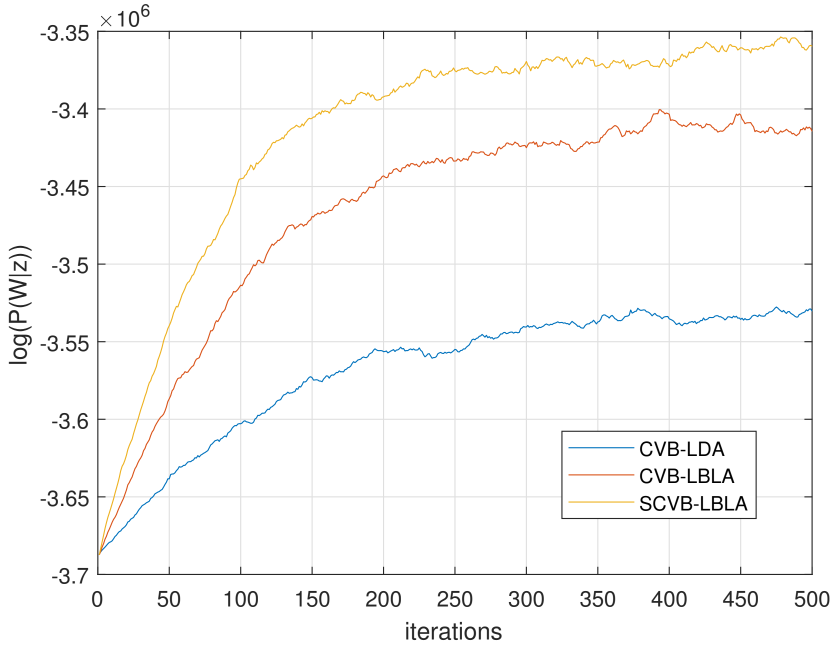


Figure 3.15: Convergence Process between the SCVB-LBLA and the batches using NIPS data

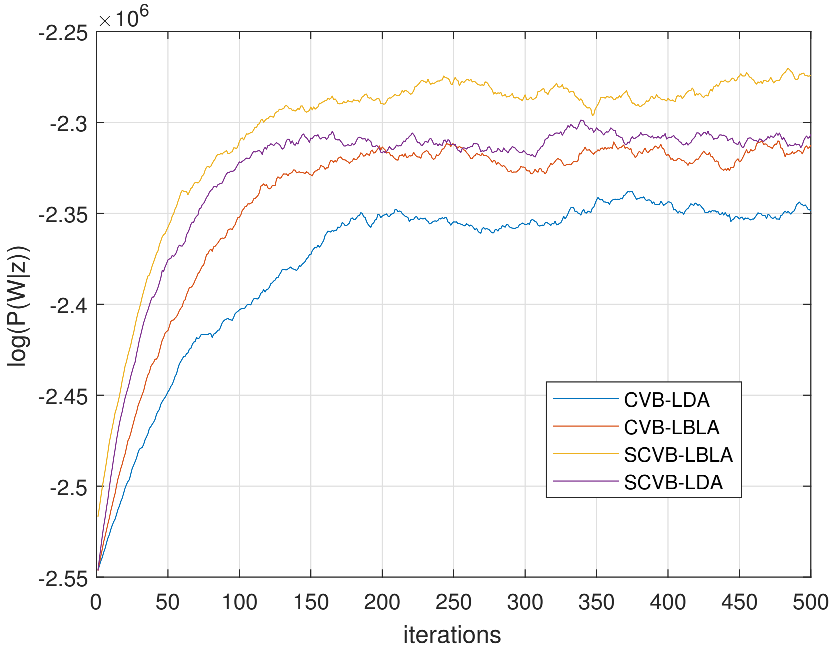


Figure 3.16: Convergence Process between the SCVB-LBLA and the batches using Natural scenes image data

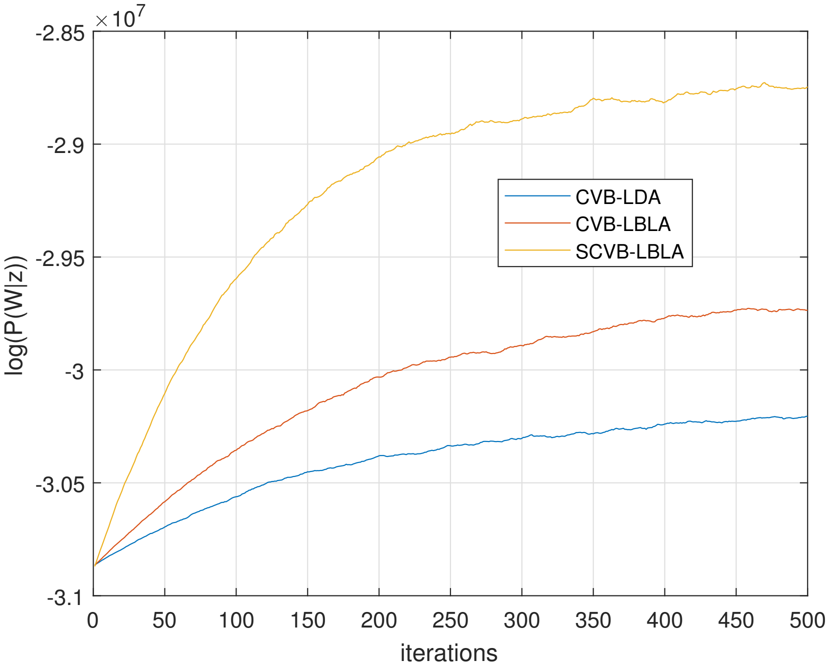


Figure 3.17: Convergence Process between the SCVB-LBLA and the batches using Activity Recognition data

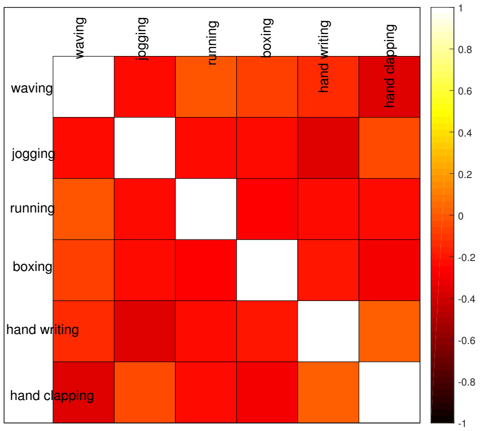


Figure 3.18: Action Recognition Dataset's correlation map

	Model	Inferences	Classes	Conjugate prior	Accuracy
Fei Fei et al. in 2005 in [2]	LDA (batch)	VB	13	Dirichlet: the prior despite its performance has been observed to have problems in intra class variation problems and for positvely correlated data	76%
Bakhtiari and Bouguila in 2014 in [64]	LBLA (online)	VB	7	Beta-Liouville (BL), flexible conjugate prior	64.30%
Our proposed work	SCVB-LBLA (stochastic)	CVB: this is the current state-of-the-art inference in topic model, as it combines the advantages of both the VB and CGS [12, 13, 67, 80]	15	Beta-Liouville, flexible conjugate prior	76.8%

Table 3.5: Our SCVB-LBLA model and other competitors performances using the same natural scene categorization dataset

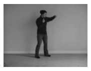
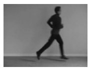
boxing	hand clapping	jogging	hand waving	Running	walking
					

Table 3.6: KTH Action Recognition Dataset

Chapter 4

Efficient Integration of Generative Topic Models Into Discriminative Classifiers Using Robust Probabilistic Kernels

We propose an alternative to the generative classifier that usually models both the class conditionals and class priors separately, and then uses the Bayes theorem to compute the posterior distribution of classes given the training set as a decision boundary. Because SVM (support vector machine) is not a probabilistic framework, it is really difficult to implement a direct posterior distribution-based discriminative classifier. As SVM lacks in full Bayesian analysis, we propose a hybrid (generative-discriminative) technique where the generative topic features from a Bayesian learning are fed to the SVM. The standard LDA (latent Dirichlet allocation) topic model with its Dirichlet (Dir) prior could be defined as Dir-Dir topic model to characterize the Dirichlet placed on the document and corpus parameters. With very flexible conjugate priors to the multinomials such as GD (generalized-Dirichlet) and BL (Beta-Liouville) in our proposed approach, we define two new topic models: the BL-GD and GD-BL. We take advantage of the geometric interpretation of our generative topic (latent) models that associate a K-dimensional manifold (K is the size of the topics) embedded into a V-dimensional feature space (word simplex) where V is the vocabulary size. Under this structure, the low dimensional topic simplex (the subspace) defines a document as a single point on its manifold and associates each document with a single probability. The SVM, with its kernel trick, performs on these documents probabilities in classification where it utilizes the maximum marging learning approach as a decision boundary. The key note is that points or documents that are close to each other on the manifold must belong to the same class. Experimental results with text documents and images show the merits of the proposed framework.

4.1 Introduction

Machine learning and AI (artificial intelligence) have been responsible for a wide variety of applications such as object detection and recognition, information retrieval, and natural language understanding and processing. These are very hot topics in the research community. Though, object categorization has always received a particular attention from researchers in the area of computer vision due to the emergence of multimedia datasets (texts, images, videos, sounds, etc) as they are increasingly becoming very complex and difficult to handle. Building models that could fully represent or describe the intrinsic characteristics in these collections of data while allowing easy classification has always been one of the top objectives and challenging tasks in machine learning. In general, object classification can be divided in two main groups in the literature: the generative approach and the discriminative scheme [81].

These two techniques can be formulated as follows: using for instance (for now) the variable Υ as the class label and χ as the observed data in class Υ, the discriminative approach will directly model the posterior distribution $p(\Upsilon/\chi)$ or estimate a function h such that $h(\chi) = \Upsilon$, from the observed data [82, 38, 81]. On the other hand, generative techniques will model both the prior distribution $p(\Upsilon)$ and the class conditional (likelihood function) $p(\chi/\Upsilon)$ separately, which is equivalent to modeling the joint distribution $p(\chi, \Upsilon)$ before estimating the posterior $p(\Upsilon/\chi)$ of the class given the training set using Bayes theorem as a decision boundary. [67, 80, 81, 38]. A real life analogy to these definitions would be to determine for instance, the type of music someone is currently listening (song). In this scenario, the generative approach will obviously learn about each music type (such as classical, jazz, country, electronic, etc.) before indicating to which type of music this particular song belongs. A discriminative method takes a much simpler and faster approach: it does not learn any of these music types. It will only focus on showing differences between the types of musics (similarities or dissimilarities). Consequently, discriminative techniques do not learn the very details about models of different classes while generative approaches do. Discriminative methods go directly to the point and often do not require lot of computational ressources as in the case of generative schemes. This simplicity and robustness (superior performance) in the discriminative approaches have often attracted many researchers [11, 82, 38, 81] since their asymptotic error is even lower than the one found in generative approaches [82]. However, generative schemes are still being implemented in many machine learning environments for their usefullness and popularity [11, 34, 7, 67, 80, 83, 84, 85, 86, 87, 23, 64]. This is because generative approaches (while requiring prior information [88]) learn about the additional details about their models which can be useful in a case of occlusion and missing data. Discriminative techniques on the other hand do not have such flexibility when facing missing data or occlusions. Generative techniques can compute marginals from the joint distributions. This is useful in applications such as outlier detection or novelty detection where the model detects efficiently new data that carry low probability and therefore very difficult to predict accurately [89]. Importantly, during the learning process for instance, generative approaches have ability to handle many (thousands) object categories better than discriminative classifiers [81]. Moreover, following the work in [82], generative schemes have also proved to outperform discriminative methods in a binary classification problem with small number of training samples. For instance, the SVM despite its discriminative power in classification is not a probabilistic approach, and it does not provide posterior distributions. Posterior distributions are important in Bayesian analysis because they provide the tool to make optimal decisions in machine learning (for instance when combining models, mimimizing risk, determining a rejection criteria that minimizes misclassification rate, etc. [89]). Therefore, their abscence makes it difficult to implement a Bayesian learning in SVM. In contrast, generative schemes benefit from a Bayesian analysis. These characteristics illustrate the strengths and capabilities of each approach. As they carry complementary advantages, it has been suggested to merge the two methods, so that their integration guarantees improvement in performance in automatic object classification. It led to the emergence of hybrid (generative-discriminative) models [11, 90, 91, 92]. Particularly, for SVM, as today's machine learning techniques carry a strong emphasis on Bayesian paradigm, combining generative models with the SVM classifier remains an essentiel step to allow this classifier to implicitly take advantage of the Bayesian learning. This has been the work of researchers such as [11] who successfully showed the flexibility of the hybrid generative-discriminative with mixtures models where the discriminative classifier is the SVM. The SVM heavily relies on efficient kernel formulation in order to provide robust classification. With the high complexity in the datasets and models, standard kernels such as linear, polynomial, Gaussian RBF (radial basis function) are very restrictive in terms of performance. Furthermore, despite the flexibility of the well-known Fisher kernel [93], it often lacks in preserving the nonlinearity induced by the generative model [94]. This is an example of the necessity to utilize appropriate kernels for better results in the hybrid, generative-discriminative models [11]. The introduction of the Fisher kernel has been immediately followed by the work of other researchers such as [95] and [96] who were able to combine generative features to SVM using the Kullback-Leibler kernel and the TOP kernel derived from Tangent vectors Of Posterior log-odds (TOP), respectively.

It is also noteworthy that recent development in the generative architecture has witnessed the emergence of topic models [97, 98, 99, 100, 101, 102] such as LDA (latent Dirichlet Allocation) [3, 2]. Originally implemented for text document modeling and analysis within the BoW (bag of words), the LDA topic model is currently dominating the area of computer vision with interesting applications related to image categorization [2], sentiment and behaviour analysis [103], text analysis through the social CQA (Community Answering Questions) platform [104], videos analysis [80], and 3D object modeling [41] for retrieval systems.

One of the successes of topic modeling is the introduction of intermediate representations within the bag of words called topics. They are low dimensional subspace representations such that documents are now described as mixtures of topics while topics are defined as distributions over the vocabulary words. This provides a hierarchical description of documents with the observed data. Though, the limitation of the Dirichlet-based topic models due to the Dirichlet (Dir) prior [23, 64, 80, 67] prompted the use of other flexible priors such as GD and BL. These conjugate priors led to some improvement in generative topic models as they provide robust inferences along with efficient generative processes [23, 64, 80, 67, 18]. In addition, the collapsed representation proposed in [13] for batch processing has shown improvement in the generative topic models implementation. However, little work has been done in the literature to connect the generative topic model to the SVM classifier to take advantage of its superior discriminative property based on maximum margin learning as a decision boundary. In the generative stage, the topic features must be generated and then in the discriminative stage, the topics are then fed to the SVM which performs the classification. This constitutes our main objective. The generative stage which learns the topics requires an efficient inference capable of delivering heterogeneous topic features.

Though, many probabilistic topic models usually implement standard variational Bayes approaches. Variational Bayes [105, 106, 23, 64, 35], despite their deterministic nature are very limited when it comes to characterize dependency betwen topic components, for instance to allow a better compression of the topic features, which is essential for the performance of our SVM classifier. In the generative stage, our proposed approach ultimately implements two robust generative topic models using asymmetric BL and GD in the collapsed space of latent variables. The superiority of the collapsed variational Bayes (CVB) inference in topic modeling is enhanced by the use of these two specific conjugate priors to the multinomials. Normally, using these two priors leads to four topic models: the BL-BL topic model, the GD-GD-topic model, the GD-BL topic model, and finally the BL-GD topic model. The first two topic models here (GD-GD and BL-BL) have been already implemented in our previous work within the CVB inference [67, 80, 107] and they represent the direct extensions to the Dir based-CVB-LDA [12]. The last two topic models (GD-BL and BL-GD) are the ones that are subjects of implementation in this chapter. Importantly, they also carry the CVB inference; and they represent the generative stage in the formation of our hybrid (generative-discriminative) model. As the generative topic features must be fed into the SVM classifier using powerful kernel functions that operate in distribution space, we therefore provide to the SVM, a collection of nonlinear probabilistic kernels (such as Jensen-Shannon kernel, symmetric Kullback-Leibler divergence kernel, Bhattacharyyaa kernel, Renyi kernel, etc.) to cope with data processing in distribution space while allowing an improved classification rate as we induce the space with the CGS (collapsed Gibbs sampler) that operates within the variational Bayesian inference[12]. It samples from the variational distribution in the collapsed space. The CVB corrects the bias in VB due to its CGS and the VB fixes the deterministic limitation of CGS [12]. Due to CVB, our generative topic features are robust, accurate and efficient [12, 13, 67, 80]. The contribution in our proposed hybrid framework is as it follows:

- With CVB inference using asymmetric GD and BL priors simultaneously, we obtained the BL-GD and GD-BL topic models that produce heterogeneous topic features in the generative stage
- SVM is not a probabilistic model; however, we successfully use the kernel trick formulation to make it operate on documents represented as topic features which are probability distributions; SVM now assigns a class label to a previously unseen document based on its topic distribution using its maximum margin framework.

Experimental results in image and text document classification show the efficiency of the proposed approach in comparison to its major competitors.

This chapter is structured as follows: section 4.2 illustrates the background and related work. Section 4.3 presents the new approach while section 4.4 covers the experiments and results in several applications. And finally, section 4.5 emphasizes on some future work and provides a conclusion.

4.2 Related work and background

In general, low performance in traditional machine learning techniques in applications such as object categorization [108, 109, 100] have led to the emergence of hybrid models especially generative-discriminative methods. This type of hybrid framework is often a combination of two stages: the generative stage which produces the features, and the discriminative stage which performs the classification using the features produced by the generative stage [11]. It is noteworthy that the complexity and characteristics in data representation often dictate the model to implement. For instance, in the past, Gaussian data dominated model learning; however, recently, the emergence of mmultimedia data causes many processing systems to work with count data especially text documents [3, 2, 13, 80, 67, 34, 50, 11]. Using the same analogy to modeling techniques, we can observe that in machine learning literature, generative models such as GMM (Gaussian Mixture Models) and HMM (Hidden Markov Models) were very specific to Gaussian data. Despite their strong assumption on parameters (as parametric distributions), these models have often received a lot of attention in the research community because of their simplicity in learning and estimation; most importantly as their functionalities were very well understood in data science [110]. So, the recent proliferation of count data led to the introduction of other generative models such as Beta-Liouville mixtures, generalized Dirichlet mixtures [7, 80, 67], Dirichlet process mixtures [39, 34, 62], and finally topic models considered as a new class of generative approaches [3, 2, 18, 64, 35, 67, 80].

Two main groups define topic models [98, 111] in the literature. We have probabilistic models (PLSA (probabilistic latent semantic analysis) and LDA) and non-probabilistic topic models such as latent semantic analysis (LSA), matrix factorization, and non-negative matrix factorization (NNMF) [112, 97]. The early success of probabilistic models especially LDA has led to other extensions to enhance the flexibility of LDA. They represent LDA-based topic models. Methods such as Patchinko Allocation topic model [20], correlated topic model [113, 114, 8], supervised topic model [115, 92, 116, 117, 118], dynamic topic model [119, 101, 120, 121], hierarchical topic model [122], spherical topic model [123], all characterize these alternatives provided to the LDA architecture. Currently, within the framework of LDA-based topic models, the advancement of social media platforms [124] and online services such as Q&A (questions and answers)[104] communities are having some serious impacts on extensions such as dynamic topic model [125, 126, 124, 101, 120], correlated topic model [114, 113], supervised topic model, and online topic model schemes [127, 128, 129, 130, 107]. Current topic models also provide improvement in semantic analysis [102, 109, 131, 132, 101] to enhance coherence in the topics estimated and the relationship between documents [133]. Some current hot topics in research (within topic modeling framework) include social network analysis, bioinformatics [134], emotion, sentiment analysis [125, 135, 126], and information retrieval [136, 41]. It is important to notice that the generative setting, through the BoW representation including its derivates and topic models, have provided tremendeous success in computer vision for object learning and categorization [2, 137, 138, 139, 67, 80]. Typical to generative techniques, probabilistic topic models use extensively prior information with distributions such as Dirichlet, Beta-Liouville, and generalized Dirichlet [81, 3, 12, 36, 140, 34, 7].

Particularly, the immediate success of the well-known topic models such as PLSI (probabilistic latent semantic indexing) or PLSA(probability latent semantic analysis) [33] and LDA in text document processing and analysis has been well received in the research community; especially, with the tremendous contributions of LDA in both text and visual document annotation and categorization [141]. As a parametric model and a generarative probabilistic technique initially implemented for topic discovery in large document collections [142], LDA [3] characterizes documents as mixtures of topics while the topics are themselves mixtures over the vocabulary words. By observing the LDA architecture, we can conclude that a very important attribute of topic models (PLSI [33] and HDP (hierarchical Dirichlet process) [14]) is their ability to operate on distribution space where their topic structures (latent variables) are defined as distributions summarizing the characteristics of the dosuments. They produce multinomial distributions over the topics given the data.

There has been a huge interest in providing extensions within the generative topic model framework by utilizing the flexibility of operating in distribution space. For instance, the work of [143] successfully builds a nonparametric topic model by replacing the document multinomial mixture model in LDA with the kernel density estimator. It is a way of solving the discretization problems related to the clustering and quantization processes during the codebook formation in topic model. It provides a framework that implicitly works on continuous feature space rather than discrete features space in topic modeling. Furthermore, authors in [41] propose a multitopic model with a model selection criteria that solves the problem of predefining a fixed number of topics for 3D object retrieval using the Kullback-Leibler divergence between 3D objects distributions within the BoW. Therefore, improving the characteristics of generative topic models coupled with the possibilities offered by working with distributions became subjects of discussions in the research community as well for tasks such as classification. The motivations include the possibility of carrying potential properties of generative topic models into discriminative classifiers to boost classification performance. This is because a recent development in discriminative setting through kernels formulation also allows SVM to perform with input features that are fully represented as distributions [144]. As a result, due to the success of LDA, recent works in machine learning and computer vision are able to provide extensions that combine LDA with discriminative classifiers [144, 145, 141]. For instance, authors in [144] provide a way of extracting latent features from probabilistic topic models in distribution space. The features are then used by the SVM for classification. Their topic model (LDA) is implemented within a Bayesian nonparametric setting using HDP (hierarchical Dirichlet process) for model selection. It leads to a topic model kernel that is robust for classification with the SVM. The work in [145] implements kernel topic models where it provides an extension to topic models by replacing the document mixture weights with Gaussian distributions leading to a Bayesian inference based on latent Gaussian. As a Gaussian process latent variable model, the technique is a combination of Gaussian process regression and LDA topic model in a nonparametric setting. In addition, authors in [141] were able to successfully perform classification on high spatial resolution remote sensing images using the LDA topic model with a kernel-based SVM that utilizes a combination of RBF or Gaussian kernels. In [67, 80], the authors provide alternatives to the LDA topic model [3] and LGDA (latent generalized Dirichlet allocation) [23] in a classification framework where they combine unsupervised learning (for the topics estimation) to a supervised technique by implementing generative classifiers for the topics similar to the work in [2]. An online version of the Naive Bayes classifier has been proposed within topic modeling environment by the same authors in [107]. In supervised learning, there have been extensive works using hybrid (generative-discriminative) models. Hybrids in general are able to demonstrate that the performance in discriminative models using SVM always depends on the characteristic of the generative features (data) and the choice of the kernels used [81, 110]. Standard kernels such as Gaussian, linear, polynomial were heavily utilized in the past in classification problems with success. This is the case of hybrids that implement for instance GMM or HMM into discriminative classifiers (SVM) using standard kernels [110] with excellent results in object categorization. The complexity in today's data and models characteristics are requesting a new generation of kernels that can cope with the challenge added to the fact that there is a huge interest in working with distributions nowadays. This automatically leads to the introduction of probabilistic kernels. Their flexibility allows a better generalization of the SVM. The SMM (support measure machine)[146] and latent SMM [147] are true examples that illustrate this generalization capabilities of the SVM: they currently represent one of the state-of-the-art techniques for object classification using distributions within the BoW framework in discriminative settings. However, originally, the Fisher kernels proposed by Jaakola and Haussler [93] catalyzed the emergence of kernels for probabilistic generative models used in discriminative classifiers today. Another kernel is the TOP kernel derived from the Tangent vectors Of Posterior log-odds. These two kernels (Fisher and TOP) were successfully used in DNA (Deoxyribonucleic acid) and protein sequence analysis (classification) [96, 93].

Other recent hybrids as they exhibit the flexibility of their generative models (based on Beta-Liouville and generalized Dirichlet mixtures) in discriminative classifiers (SVM) have reported similar success in image categorization [11, 44] while using probabilistic kernels. Despite the major contributions shown by previous and some recent schemes, they still carry some limitations. For instance, as we emphasize on topic models in this chapter, the Dirichlet conjugate prior often affects LDA's performance for positively correlated data. This is because it has a very restricted covariance structure compared to GD and BL that are more flexible [23, 35, 80, 67]. Many topic models in the literature are LDA-based. This could have a negative impact on the generative process and inferences [67, 80] in Dirichlet-based topic models such as LDA. In addition, the possibility of using topic models in discriminative classifiers has created many extensions within the nonparametric setting to account for efficient model selection and processing. However, working with nonparametric models could be very challenging as they require operation or modeling in infinite dimensional spaces. For instance, in [145], the kernel topic model implemented is a Gaussian process latent variable model based on LDA. It has a very complex inference as the framework is not analytically tractable. Similar challenges are noticed in inferences in the work of [144] with the implementation of the HDP in LDA model as it sets the number of latent factors or topics into infinite. Furthermore, the SMM and latent SMM [146, 147] have also provided insights on the possibilty of using the concept of distributions within the discriminative platform itself. They have good mathematical foundations and formulations about the space that could allow such implementations as they apply their work in the RKHS (reproducing kernel Hilbert space) that is equipped with an embedding kernel and inner product; however, these techniques in overall could be very complex and require knowledge of vector spaces such as Hilbert spaces which are generalizations of the Euclidean space in finite or infinite dimension. These methods are not hybrids of the type generative-discriminative. They are dedicated discriminative classifiers working indirectly (implicitly) on distributions by using standard kernels in the RKHS [146, 147] where the probability distributions are represented as mean embeddings [146]. Because these methods operate on standard kernels, nonlinear probabilistic kernels such as symmetric KL (Kullback-Leibler) divergence could not be defined directly on the RKHS because of the inner product operation on the Hilbert space.

As we consider all the different characteristics within previous methods that include generative models, (especially topic models), discriminative approaches using SVM, kernels, and hybrid (generative-discriminative) techniques, we propose, in this chapter, an extension in topic modeling framework using finite mixtures, similar to LDA. We especially implement a new approach (hybrid method) that integrates the flexibility of our generative topic model into a powerful discriminative classifier (SVM). It is equipped with well-defined nonlinear probabilistic kernels that allow analysis in distribution space using empirical likelihood (EL) framework for generative topic models in SVM. Within our proposed approach, the use of EL provides distribution estimations. Importantly, with a combination of two different priors (asymmetric GD and BL) used simultaneously within the same generative process, our proposed method introduces a collapsed representation through the collapsed variational Bayesian inference that allows estimation of exact posteriors and easy access to convergence. Most previous generative topic models are either variational-based inference [148] techniques (provide convergence, but posterior estimations are often not exact [105, 106, 12]) or collapsed Gibbs sampling-based methods (posterior distributions estimations are exact; however, they suffer from convergence [12]). In contrast, our proposed generative topic model is obtained from a combination of two inferences: VB and CGS. It follows the work in [12] which introduced the CVB (collapsed variational Bayesian inference) for LDA. It is one of the state-of-the-art inferences in topic modeling. Though, because of the limitations of the Dirichlet distribution in LDA [23, 64, 35], we provide alternatives with the use of Beta-Liouville and generalized Dirichlet conjugate priors.

The generative topic model in our proposed approach, because, based on LDA, automatically introduces hierarchies in the observed data with the use of topics as intermediate representations. So, the topic representation in our proposed method could be for instance an alternative to generative models using Beta-Liouville mixtures and generalized Dirichlet mixtures [7]. Using our proposed framework with nonlinear probabilistic kernels, we obtain a system that finally gives us tools to represent any object or document as a distribution parameterized by two mean variables: the document-topic parameter and the topic-word parameter.

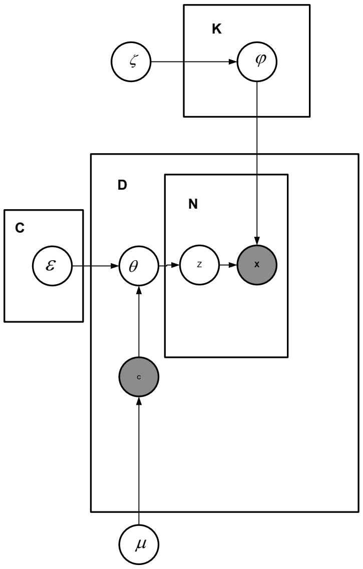


Figure 4.1: Generative stage using topic (latent) graphical model. The shaded circle denotes observed variables $\boldsymbol{x}$ and the class c

.

4.3 Proposed Approach

We implement a classification framework where the classifier, the SVM, gets its features from our generative topic models which simultaneously use asymmetric BL and GD as conjugate priors to the multinomial distributions. Documents (images, texts) are first represented as distributions using characteristics of our generative topic models, and then they are presented to the support vector machine for classification. This setting ultimately constitutes our generative-discriminative method that utilizes nonlinear probabilistic kernels. The generative topic models implemented in this chapter follow the graphical representation previously proposed in [80, 67, 2, 107] for object classification using intermediate representations such as topics [2] as shown in Fig. 4.1. Based on the LDA architecture [3], the extensions (generative topic models) we are also providing in this chapter are a result of sampling documents and corpus parameters using asymmetric GD and BL priors, simultaneously. In this scenario, the proposed generative process uniquely offers the possibility to either 1) draw the documents parameters from the BL while the corpus parameters are sampled from GD or 2) sample the documents parameters from GD while the corpus parameters are drawn from the BL distribution. This leads to the implementation of two topic models in our proposed generative framework.

4.3.0.1 Research objectives

Many techniques related to classification using the hybrids, generative topic models-discriminative methods do not always fulfill the following requirements: 1) The flexibility and the structure (symmetric or asymmetric) of the prior 2) the robustness of the generative process including inference techniques and 3) the choice of kernels. In a supervised topic modeling, these characteristics and requirements are intimately related to each other [9]. However, many hybrid techniques using topic models are just partially robust because they lack some of these essential requirements. In our proposed method, we are mainly implementing a system of integration that takes into account each of these requirements where we provide a combination of much capable and flexible priors (than the Dirichlet) that first helps improving the generative process and inferences.

A much improved inference technique is essential for an accurate parameter estimation that increases the coherence and robustness of our generative features and kernel functions formulation. This is the essence of our hybrid model as we formulate a complete framework where we combine two different and flexible priors (BL and GD) within the collapsed variational inference that enables robust generative features for our kernel machine. In addition, the flexibility of our priors and inferences allow us to handle with efficiency inter and intraclass variation problems due to the ability of our method to deal with correlation and semantic analysis effectively. And this includes the possibility of working with a variety of datasets.

Our proposed method in its hybrid setting guarantees the best generative topic model and the best discriminative method as we also believe that the SVM is the appropriate candidate in large scale processing compared to the standard Naive Bayes classifier widely used in classification framework that implements topic models [2].

4.3.0.2 Beta-Liouville and generalized Dirichlet distributions

The generalized Dirichlet (GD) distribution was already introduced and defined in [80, 67, 64]. In this chapter, we also present the Beta-Liouville (BL) distribution (another flexible conjugate prior with a more versatile covariance structure) [35, 64, 11, 7]. Compared to LDA [3], both priors (GD and BL) are now replacing the Dirichlet distribution in topic modeling. We also emphasize on the use of asymmetric priors compared to symmetric ones as they have a direct impact on the robustness of the generative topic models [9].

In a $(K+1)$-dimensional space, the BL distribution with parameters $\varepsilon =$

$(\alpha_1, ..., \alpha_K, \alpha, \beta)$ also written as BL(ε) could be defined as:

$$p(\vec{P}|\varepsilon) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \prod_{k=1}^{K} \frac{P_k^{\alpha_k - 1}}{\Gamma(\alpha_k)} \left(\sum_{k=1}^{K} P_k\right)^{\alpha - \sum_{k=1}^{K}\alpha_k} \left(1 - \sum_{k=1}^{K} P_k\right)^{\beta - 1} \tag{124}$$

where $\vec{P} = (P_1, ..., P_K)$ is a K-dimensional random variable. Using the notion of conjugate prior to the multinomial, if $\vec{P} = (P_1, ..., P_K)$ follows a Beta-Liouville distribution with parameter θ while the vector of counts $\vec{X_i} = (X_1, ..., X_{D+1})$ is drawn from a multinomial distribution with parameter $\vec{P}$, then the posterior distribution $p(\vec{P}|\varepsilon, \vec{X_i})$ is also a Beta-Liouville. It therefore leads to the following updates in the posterior distribution $p(\vec{P}|\varepsilon, \vec{X_i})$.

$$\begin{cases} \alpha'_k = \alpha_k + X_k \\ \alpha' = \alpha + \sum_{k=1}^{K} X_k \\ \beta' = \beta + X_{K+1} \end{cases} \tag{125}$$

As previously mentioned, the implementation of our proposed approach using two conjugate and asymmetric priors (BL and GD) simultaneously, leads to two generative topic models: the first model draws the document parameter from GD while the corpus is sampled from the BL. In addition, it uses the collapsed variational inference (CVB), that is one of the state-of-the-art inference techniques in topic modeling [80, 67, 12]. We call it the CVB-GD-BL-based topic model or *topic model I*. On the other hand, similarly, the second method uses BL for the document parameter and GD for the corpus parameter within the CVB inference leading to CVB-BL-GD-based topic model. This is *topic model II*.

4.3.0.3 Generative Processes

LDA is recognized as the simplest topic model where each document is is a mixture of K topics in different proportions. Documents while being maintaing K topics in different proportions must belong to same class. This to characterize the observe data. Though in our proposed approach, we the BL and GD priors replace the Dir distribution. For instance, in the GD-BLtopic model, the generative process is now expressed as follows:
1-We draw topics from $\varphi_k \in$ BL(ζ) for $k \in \{1, 2, 3, ..., K\}$ where $\zeta = (\lambda_{kv}, ..., \lambda_{kV}, \lambda, \eta)$
2-We draw each document $j \in \{1, ..., D\}$
(a) Draw topic proportions $\theta^c \sim GD(\varepsilon)$
where $\varepsilon = (\alpha_{c1}, \beta_{c1}, ..., \alpha_{cK}, \beta_{cK})$ and $c \in \{1, 2, ..., C\}$
(b) For each word $x \in \{1, ..., N\}$
i) Draw topic assignments
$z_{jn} \sim$ Multinomial (θ_d^c)
ii) Draw word
$x_{jn}|z_{jn}, \varphi_k \sim$ Multinomial $(\varphi_{k z_{jn}})$
We could therefore provide a generative of the BL-GD-topic model as well following the same scheme.

4.3.1 CVB-GD-BL-based topic model

Using concepts such as patches for images [80, 67] (similar to words for text analysis) within the BoW, we implicitly elaborate on document representation as visual features in topic modeling framework. In contrast to the standard Naive Bayes classifier for topic modeling, we simply implement in our proposed approach an improved supervised topic model that uses SVM in single-label classification problems. One major contribution is that our proposed method is ultimately done with (a combination of) better priors that provide much flexible generative processes leading to robust inferences and generative features for our kernel functions formulation. In this framework, we can use the variable $\mathcal{X}$ and W interchangeably to denote the collection of words or patches (visual words) in a document or object within the BoW.

4.3.1.1 Bayesian inference using asymmetric GD and BL priors

From the work presented in [80, 67], the generative equation in the fully collapsed space is given by:

$$p(\mathcal{X}, z|\varepsilon, \zeta, c) = \int_{\theta}\int_{\varphi} p(\mathcal{X}, z, \theta, \varphi|\varepsilon, \zeta) d\varphi d\theta \tag{126}$$

Due to the prior conjugacy between both GD and BL with respect to the multinomial distribution, Eq. 126 becomes easy to compute as it is now expressed as a product of Gamma functions. As a result, the generative equation of the proposed model in the collapsed space of latent variables is:

$$\begin{aligned}p(\mathcal{X}, z|\varepsilon, \zeta, c) = \prod_{j=1}^{D}\left[\prod_{i=1}^{K}\frac{\Gamma\left(\alpha_{ci}+\beta_{ci}\right)}{\Gamma\left(\alpha_{ci}\right)\Gamma\left(\beta_{i}\right)}\right] \\ \times\left[\prod_{i=1}^{K}\frac{\Gamma\left(\alpha'_{ci}\right)\Gamma\left(\beta'_{ci}\right)}{\Gamma\left(\alpha'_{ci}+\beta'_{ci}\right)}\right]\left[\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{r=1}^{V}\lambda_r\right)\Gamma\left(\lambda+\eta\right)}{\Gamma\left(\lambda\right)\Gamma\left(\eta\right)\prod_{r=1}^{V}\Gamma\left(\lambda_r\right)}\right] \\ \times\left[\frac{\Gamma\left(\lambda'\right)\Gamma\left(\eta'\right)\prod_{r=1}^{V}\Gamma\left(\lambda'_r\right)}{\Gamma\left(\sum_{r=1}^{V}\lambda'_r\right)\Gamma\left(\lambda'+\eta'\right)}\right]\end{aligned} \tag{127}$$

The equation provided by the joint $p(\mathcal{X}, z|c, \varepsilon, \zeta)$ finally shows some updates due to the multinomial distributions. In the document-topic update in class c, we have:

$$\alpha'_{ci} = \alpha_{ci} + N^{i}_{j(.)} \qquad\qquad \beta'_{ci} = \beta_{ci} + \sum_{l=i+1}^{K+1} N^{l}_{j(.)} \tag{128}$$

In the topic-word update, it shows:

$$\begin{cases}\lambda'_r = \lambda_r + N^{i}_{(.),r}\\ \lambda' = \lambda + \sum_{r=1}^{V} N^{i}_{(.),r}\\ \eta' = \eta + N^{i}_{(.),V+1}\end{cases} \tag{129}$$

From this point, performing a Bayesian inference in the fully collapsed space is equivalent to approximating the conditional distribution of the latent variable $p(z|\mathcal{X}, \varepsilon, \zeta)$. By integrating out the parameters, the collapsed Gibbs sampler's equation is obtained as an expectation expression:

$$\begin{aligned}p(z_{ij} = k|\mathcal{X}, \varepsilon, \zeta, c) = \\ E_{p(z^{-ij}|\mathcal{X},\varepsilon,\zeta,c)}[p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c)]\end{aligned} \tag{130}$$

such that:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) \propto \left[\frac{(N_{jk.}^{-ij} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N_{jl.}^{-ij})}\right] \times \left[\frac{(\lambda + \sum_{r=1}^{V} N_{.kr_{ij}}^{-ij})}{(\lambda + \eta + \sum_{r=1}^{V+1} N_{.kr_{ij}}^{-ij})}\right] \times \left[\frac{(\lambda_v + N_{.kv_{ij}}^{-ij})(\eta + N_{.k(V+1)_{ij}}^{-ij})}{(\sum_{r=1}^{V} N_{.kr_{ij}}^{-ij} + \lambda_r)}\right] \tag{131}$$

Normalizing the distribution above leads to a posterior probability defined as:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) = \frac{A(k)}{\sum_{k'=1}^{K} A(k')} \tag{132}$$

such that:

$$A(k) = \left[\frac{(N_{jk.}^{-ij} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{(\alpha_{k} + \beta_{ck} + \sum_{l=k}^{K+1} N_{jl.}^{-ij})}\right] \times \left[\frac{(\lambda + \sum_{r=1}^{V} N_{.kr_{ij}}^{-ij})}{(\lambda + \eta + \sum_{r=1}^{V+1} N_{.kr_{ij}}^{-ij})}\right] \times \left[\frac{(\lambda_v + N_{.kv_{ij}}^{-ij})(\eta + N_{.k(V+1)_{ij}}^{-ij})}{(\sum_{r=1}^{V} N_{.kr_{ij}}^{-ij} + \lambda_r)}\right] \tag{133}$$

4.3.1.2 CVB inference with asymmetric priors

In general, the main goal in Bayesian inference is the estimation of models hidden variables (models parameters and latent variables). This is equivalent to computing the joint posterior distribution $p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c)$. Though, the posterior distribution in topic modeling framework is often intractable because the denominator of the posterior equation, the normalizing factor, is not tractable. This normalizing factor is the marginal likelihood. Therefore, inference techniques such as VB and CGS from MCMC (Markov chain Monte Carlo) are often used for hidden variables estimations.

The collapsed varitational Bayesian inference implemented in our proposed approach is essentially a VB in the collapsed space of latent variables induced by the CGS (Eqs. 130 to 133). As usual, performing VB inference is equivalent to introducing a set of variational distributions (exponentional family) $\hat{Q}(z, \theta, \varphi)$ that minimize the Kullback-Leibler divergence (KL) between the joint variational distribution $\hat{Q}(z, \theta, \varphi)$ and the true joint posterior distribution $p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c)$. The scheme also introduces a lower bound (evidence lower bound or ELBO) to the log marginal likelihood $\log p(\mathcal{X}|\varepsilon, \zeta, c)$. And maximizing the ELBO is equivalent to minimizing the $\mathrm{KL}(\hat{Q}(z, \theta, \varphi)||\ p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c))$. The lower bound (ELBO) to the log marginal likelihood can be considered as an upper bound (negative ELBO) to the negative log marginal likelihood. So instead of maximizing the ELBO, we could minimize the negative ELBO. This negative ELBO is a functional

acting on the joint variational posterior distribution following the work in [12]. It is called variational free energy ($\tilde{\mathscr{F}}(\tilde{Q})$) in the joint space and $\hat{\mathscr{F}}(\hat{Q})$ in the collapsed space).
In CVB, minimizing the variational free energy with respect to $\hat{Q}(\theta, \varphi|z)$ and then with respect to $\hat{Q}(z_{ij}|\hat{\psi}_{ij})$ leads to $\hat{\mathscr{F}}(\hat{Q}(z))$ such that:

$$\begin{aligned}\hat{\mathscr{F}}(\hat{Q}(z)) \triangleq \min_{\hat{Q}(\theta,\varphi|z)} \hat{\mathscr{F}}(\hat{Q}(z)\hat{Q}(\theta,\varphi|z)) = \\ \mathbb{E}_{\hat{Q}(z)}[-\log p(\mathcal{X}, z|\varepsilon, \zeta)] - \mathscr{H}(\hat{Q}(z))\end{aligned} \tag{134}$$

Following the work in [80, 67, 107, 13, 12] and using Eqs. 131, 132, and 133, the minimization of the functional $\hat{\mathscr{F}}(\hat{Q}(z))$ in Eq. 134 with respect to the variational distribution $\hat{\psi}_{ijk}$ finally gives the following CVB update equation using the Gaussian approximation:

$$\begin{aligned}\hat{\psi}_{ijk} = \hat{Q}(z_{ij} = k) \propto \Big(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}]\Big) \\ \times \Big(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}]\Big) \\ \times \left(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}^{-ij}]\right) \\ \times \Big(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}^{-ij}]\Big) \\ \times \Big(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}^{-ij}]\Big) \\ \times \left(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}]\right)^{-1} \\ \times \left(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}^{-ij}]\right)^{-1} \\ \times \left(\sum_{r=1}^{V} \lambda_r + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}]\right)^{-1} \times \mathbb{G}\end{aligned} \tag{135}$$

such that:

$$
\begin{aligned}
\mathbb{G} = \exp\left(-\frac{Var_{\hat{Q}}(N_{jk.}^{-ij})}{2(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.k.}^{-ij})}{2(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{l=k}^{K+1} N_{jl.}^{-ij})}{2(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.kx_{ij}}^{-ij})}{2(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kxij}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.k(V+1)_{ij}}^{-ij})}{2(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{l=k}^{K+1} N_{jk.}^{-ij})}{2(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{r=1}^{V+1} N_{.kr_{ij}}^{-ij})}{2(\eta + \lambda + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(N_{.k.}^{-ij})}{2(\mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}] + \sum_{r=1}^{V} \mathbb{E}_{\hat{Q}}[\lambda_r])^2}\right) \quad (136)
\end{aligned}
$$

where:
$\mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{i'jk}$; $\mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{ijk}(1 - \hat{\psi}_{i'jk})$ in a class. The superscript $-ij$ means all the words except the word ij. It is important to notice that this update equation in CVB is the result from implementing a topic model (CVB-GD-BL-based topic model) where the document and corpus parameters are drawn from asymmetric GD and BL, respectively.

4.3.1.3 Predictive distributions from the CVB-based topic model

After the sampling process reaches a stationary distribution (convergence), the model parameters that have been initially marginalized out in the fully collapsed space are now estimated. For large samples [13, 12], the document predictive distribution in our proposed CVB-GD-BL topic model is therefore given by:

$$\hat{\theta}_{jk}^{c} = \frac{\left(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}]\right)\left(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}]\right)}{\left(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}]\right)} \quad (137)$$

Conditional on the topic k, the predictive distribution of the words φ_{kv} is:

$$\hat{\varphi}_{kv} = \left(\frac{(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}])(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}])}{(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}])}\right)\left(\frac{(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}])}{(\mathbb{E}_{\hat{Q}}[N_{.k.}] + \sum_{r=1}^{V} \lambda_r)}\right) \quad (138)$$

Following estimation of the predictive distributions (model parameters), the empirical log likelihood could be computed since it is defined as:

$$p(X_j|\theta_j^c, \varphi, \varepsilon, \zeta) = \prod_{ij} \sum_{k} \hat{\theta}_{jk}^c \hat{\varphi}_{kx} \tag{139}$$

following the work in [42], [12], [18] such that the following expected counts, $\mathbb{E}_{\hat{Q}}[N_{j..}]$, $\mathbb{E}_{\hat{Q}}[N_{.k.}]$, $\mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}]$, $\mathbb{E}_{\hat{Q}}[N_{jk.}]$, $\mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}]$, and $\mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}]$ of the unseen document are obtained from the CVB-GD-BL sampling process. The parameters of the unseen document are then used to predict its likelihood. The EL implemented in this chapter ultimately follows the work in [18, 12, 80].

4.3.2 The CVB-BL-GD-based topic model

Using the framework in [80, 67, 107, 12] and the derivations obtained from subsection 4.3.1 in this chapter, the generative equation in the collapsed space (for M documents and K topics) in our second proposed topic model is:

$$\begin{aligned} p(\mathcal{X}, z|\varepsilon, \zeta, c) = \prod_{j=1}^{M} & \left[\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_{ci}\right) \Gamma\left(\alpha_c + \beta_c\right)}{\Gamma\left(\alpha\right) \Gamma\left(\beta_c\right) \prod_{i=1}^{K} \Gamma\left(\alpha_{ci}\right)} \right] \\ & \times \left[\frac{\Gamma\left(\alpha_c'\right) \Gamma\left(\beta_c'\right) \prod_{i=1}^{K} \Gamma\left(\alpha_i'\right)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i'\right) \Gamma\left(\alpha_c' + \beta_c'\right)} \right] \\ & \times \prod_{j=1}^{M} \left[\prod_{i=1}^{K} \frac{\Gamma\left(\lambda_r + \eta_r\right)}{\Gamma\left(\lambda_r\right) \Gamma\left(\eta_r\right)} \prod_{i=1}^{K} \frac{\Gamma\left(\lambda_r'\right) \Gamma\left(\eta_r'\right)}{\Gamma\left(\lambda_r' + \eta_r'\right)} \right] \end{aligned} \tag{140}$$

where the document-topic update in a class is:

$$\begin{cases} \alpha_{ci}' = \alpha_{ci} + N_{j,(.)}^i \\ \alpha_c' = \alpha_c + \sum_{i=1}^{K} N_{j,(.)}^i \\ \beta_c' = \beta_c + N_{j,(.)}^{K+1} \end{cases} \tag{141}$$

The topic-word update is:

$$\lambda_r' = \lambda_r + N_{(.),r}^i \qquad\qquad \eta_r' = \eta_r + \sum_{d=v+1}^{V+1} N_{(.)d}^i \tag{142}$$

In the collapsed space, as we integrate out the parameters, the collapsed Gibbs sampler's equation is computed as follows:

$$\begin{aligned} p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) \propto & \left[(\alpha_{ck} + N_{jk.})\right] \\ & \times \left[\frac{(\lambda_v + N_{.kv_{ij}})(\eta_v + \sum_{d=v+1}^{V+1} N_{.kd_{ij}})}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} N_{.kd_{ij}})} \right] \end{aligned} \tag{143}$$

So, normalizing the distribution above provides a posterior probability defined as:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta) = \frac{\mathcal{A}(k)}{\sum_{k'=1}^{K} \mathcal{A}(k')} \tag{144}$$

such that:

$$\mathcal{A}(k) = (\alpha_{ck} + N_{jk.}) \frac{(\lambda_v + N_{.kv_{ij}})(\eta_v + \sum_{d=v+1}^{V+1} N_{.kd_{ij}})}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} N_{.kd_{ij}})} \tag{145}$$

Following similar steps in subsection 4.3.1, we reach the final variational update for the CVB-based framework in the second generative topic model where we use the BL and GD for document and corpus parameters, respectively:

$$\begin{aligned}
\hat{\psi}_{ijk} = \hat{Q}(z_{ij} = k) \propto& \\
&\Bigg\{ \Big[\Big(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}] \Big) \Big] \\
&\times \left[\frac{\Big(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}] \Big) \Big(\eta_\nu + \sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}] \Big)}{\Big(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}] \Big)} \right] \\
&\times \exp \left(- \frac{Var_{\hat{Q}} \left(N_{jk.}^{-ij} \right)}{2(\alpha_k + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2} \right) \\
&\times \exp \left(- \frac{Var_{\hat{Q}} \left(N_{.k\nu_{ij}}^{-ij} \right)}{2(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}])^2} \right) \\
&\times \exp \left(- \frac{Var_{\hat{Q}} \left(\sum_{d=\nu+1}^{V+1} N_{.kd_{ij}}^{-ij} \right)}{2 \left(\eta_\nu + \sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}] \right)^2} \right) \\
&\times \exp \left(\frac{Var_{\hat{Q}} \left(\sum_{d=\nu}^{V+1} N_{.kd_{ij}}^{-ij} \right)}{2 \left(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}^{-ij}] \right)^2} \right) \Bigg\}
\end{aligned} \tag{146}$$

The parameters estimates for the topic model is as follows:

$$\hat{\theta}_{jk}^{c} = \frac{\Big(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}] \Big)}{\Big(\mathbb{E}_{\hat{Q}}[N_{j..}] + \sum_{i=1}^{K} \alpha_{ci} \Big)} \tag{147}$$

The predictive distribution of the words φ_{kw} is:

$$\hat{\varphi}_{kv} = \frac{\Big(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kv_{ij}}] \Big) \Big(\eta_v + \sum_{d=v+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}] \Big)}{\Big(\lambda_v + \eta_v + \sum_{d=v}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}] \Big)} \tag{148}$$

4.3.3 Discriminative framework: SVM, kernels, and discrete distributions

A probability kernel is defined as the mapping $\mathscr{K} : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ with $\mathscr{P}$ defined as the space of probability distributions [149].

For instance, let $\mathcal{X}_i = \{x_{i1}, x_{i2}, ..., x_{iM}\}$ and $\mathcal{X}_j = \{x_{j1}, x_{j2}, ..., x_{jM}\}$ be two sequences of vectors for two multimedia objects X_i and X_j, respectively. Then, each object is associated with its probablility density function $p(x|\Omega_i)$ and $q(x|\Omega_j)$, respectively. These are parametric distributions such that Ω_i is the parameter for object X_i while Ω_j is the parameter for object X_j. When implementing topic models especially in computer vision, the bag of visual words scheme leads to the discretization of the continuous visual features space as we perform clustering and quantization methods for the elaboration of the codebook [143, 144]. This discretization causes the reformulation of the kernels using discrete distributions instead of PDFs (probability density functions).

For our generative topic model framework in the collapsed space, we recover the parameters through sampling process of the topic assignments $\boldsymbol{z}$. Let Ω be defined as $\Omega = \{\theta^c, \varphi\}$ such that θ^c is $1 \times K$ vector (document-topic parameter) and φ is a $K \times V$ (word-topic) parameter for the corpus such that its entries are φ_{kv} from $\varphi = \{\varphi_{kv}\}$. Let $\hat{\Omega}$ be the estimate of Ω such that $\hat{\Omega} = \{\hat{\theta^c}, \hat{\varphi}\}$. With $\hat{\Omega}$, we can efficiently represent the PMF (probability mass function) of each document X_j. In other words the SVM carries the generative predictive distributions for each document obtained by marginalizing out the topic model parameters. With documents in the generative stage equipped with probabilities (Eq. 139), we can define the different probabilistic kernels in the following section. Let P and Q be two distributions defined on the space Δ such that $p(x)$ and $q(x)$ represent the densities of P and Q, respectively. For our supervised topic model framework using SVM, we replace the kernel formulation in the standard (original) feature space by the one in the distribution space that accounts for topic generative features as shown in:

$$\mathscr{K}(\mathcal{X}_i, \mathcal{X}_j) \Rightarrow \mathscr{K}(P, Q) \tag{149}$$

There have been many ways of characterizing the generative structure (features) in topic models. For instance authors in [41] in 3D object retrieval system use the LDA document topic proportions θ and the KL divergence to compute the distance between two 3D objects. In their work, the topic proportions θ represent an object. However, in [127], authors implement the Jensen-Shannon divergence by considering the topics φ_k themseleves to evaluate the change in topics between two successive time slices. Similar choice is suggested in [130] where authors define the topic as a vector of probabilities over the space of words and then formulate the KL divergence between two topic distributions to assess their dissimilarity.

As topic mixtures are parameterized by θ^c while the topics themselves are parameterized by φ, we decide in our proposed approach to parameterize each document with both θ^c and φ. This representation is in line with the definition of a document in topic modeling which is described as a mixture over topics where each topic is a mixture over the vocabulary. Therefore, each discrete document multinomial distribution fed to the SVM could be described by Eq. 139 as a PMF parameterized by θ_j^c and φ .

4.3.3.1 The Kullback-Leibler kernel

Based on information divergence measures (where the measure here is the KL divergence), this probabilistic kernel computes the dissimilarity between two probability density functions $p(x|\Omega_i)$ and $q(x|\Omega_j)$ defined on the support (space) Δ:

$$D_{KL}(P, Q) = \int_{\Delta} p(x|\Omega_i) \log\left(\frac{p(x|\Omega_i)}{q(x|\Omega_j)}\right) dx \tag{150}$$

The KL divergence between P and Q ($KL(P||Q)$) could be seen as the additional amount of bits needed to encode samples from P distribution using a Q distribution-based code [130, 89].

From the KL divergence measure, we can evaluate the symmetric KL divergence as:

$$D_{SKL}(P,Q) = \int_{\Delta} p(x|\Omega_i)\log\left(\frac{p(x|\Omega_i)}{q(x|\Omega_j)}\right)dx \\ + \int_{\Delta} q(x|\Omega_j)\log\left(\frac{q(x|\Omega_j)}{p(x|\Omega_i)}\right)dx \quad (151)$$

For discrete probability distributions $P(x)$ and $Q(x)$, we can reformulate $D_{KL}(P,Q)$ over the support Δ as:

$$D_{SKL}(P,Q) = \sum_{x\in\Delta} P(x)\log\left(\frac{P(x)}{Q(x)}\right) + \sum_{x\in\Delta} Q(x)\log\left(\frac{Q(x)}{P(x)}\right) \quad (152)$$

Once the symmetric KL divergence measure is defined, the KL kernel [110] is estimated by exponentiating the symmetric KL divergence.

$$\mathscr{K}(X_i,X_j) \Rightarrow \mathscr{K}\,(P,Q)) \Rightarrow \exp\left(D_{SKL}(P,Q)\right) \quad (153)$$

4.3.3.2 The Jensen-Shannon kernel

It is based on Jensen-Shannon (JS) divergence [150] as it measures the similarity between two distributions. The JS divergence between distributions P and Q is defined as:

$$JS(P||Q) = \mathscr{H}[\upsilon P + (1-\upsilon)Q] - \upsilon\mathscr{H}[P] - (1-\upsilon)\mathscr{H}[Q] \quad (154)$$

with υ a parameter and $\mathscr{H}[P]$ the Shannon entropy of P over the space Δ is $\mathscr{H}[P] = -\int_{\Delta} p(x|\Omega_i)\log p(x|\Omega_i)dx$ such that p is the density of distribution P. A discrete formulation of the Shannon entropy is:

$$\mathscr{H}[P] = -\sum_{x\in\Delta} P(x)\log P(x) \quad (155)$$

The Jensen-Shannon kernel is obtained by exponentiating the JS divergence.

$$\mathscr{K}_{JS}(P,Q) = \exp(-aJS(P||Q)) \quad (156)$$

The JS could also be formulated using the KL by setting $g(x) = \frac{1}{2}p(x)+\frac{1}{2}q(x)$ with $\upsilon = 1/2$

$$JS(P||Q) = \frac{1}{2}KL(P||G) + \frac{1}{2}KL(Q||G) \quad (157)$$

4.3.3.3 The Bhattacharyya kernel

It is a member of the probability product kernel (PPK) family [151] that is defined as:

$$\mathscr{K}_{\rho}(P,Q) = \sum_{x\in\Delta} P(x)^{\rho}Q(x)^{\rho} \quad (158)$$

such that ρ is a parameter. Following the formulation in Eq. 158, we can define the Bhattacharyya kernel [152] as a PPK at $\rho = \frac{1}{2}$:

$$\mathscr{K}_{\frac{1}{2}}(P,Q) = \sum_{x\in\Delta} \sqrt{P(x)}\sqrt{Q(x)} \quad (159)$$

However, when $\rho = 1$, the PPK becomes the expected likelihood kernel, also called the correlation kernel as it measures the corelation between two distributions such that:

$$\mathscr{K}_1(P,Q) = \sum_{x\in\Delta} P(x)Q(x) \tag{160}$$

Because it is related to traditional linear kernels, the correlation kernel is called probabilistic linear kernel [149].

4.3.3.4 The Renyi kernel

Straight from the Shannon entropy theory, the Renyi kernel is based on the Renyi divergence measure of order σ:

$$D_\sigma(P||Q) = \frac{1}{\sigma - 1} \log \sum_{x\in\Delta} P(x)^\sigma Q(x)^{1-\sigma} \tag{161}$$

where $\sigma > 0$ and $\sigma \neq 0$
By exponentiating the symmetric Renyi divergence, it leads to the Renyi kernel that is defined as: $\mathscr{K}_R(P,Q) = \exp\{-a(D_\sigma(p(x|\Omega_i)||q(x|\Omega_j)) + D_\sigma(q(x|\Omega_j)||p(x|\Omega_i))\}$
where $a > 0$.

$$\mathscr{K}_R(P,Q) = \left[\log \sum_{x\in\Delta} P(x)^\sigma Q(x)^{1-\sigma}\right]\left[\log \sum_{x\in\Delta} Q(x)^\sigma P(x)^{1-\sigma}\right]^{\frac{a}{1-\sigma}} \tag{162}$$

The Renyi divergence is a generalization of the KL divergence, and both are identical when $\sigma \to 1$. In addition, the Renyi kernel becomes a PPK when $a = \frac{1-\sigma}{2}$. It also a Bhattacharyya kernel for $\sigma = \frac{1}{2}$.

4.3.4 Time and memory complexities

The time and memory complexity have been presented in many topic model publications [95, 153, 18, 12, 80, 89]. Though the work of [153] provided the most extensive details about time and memory complexities when processing large collections under LDA. Following the work in [153] For D documents containing each N words from a vocabulary of size V, in a particular class c, we obtain a $D \times V$ matrix where $NN0$ is the total number of nonzero elements in this document-word (sparse) matrix. During the formation of K topics, it involves placing a $K + 1$-dimensional variational distribution on every word leading to a $K \times NN0$ matrix. The parameter estimation provided the predictive document-topic distribution θ_j^c of size $K \times D$ and the topic-word predictive distribution of size $K \times V$. CVB-LDA carries $K \times NN0$ matrix along with two copies $th\hat{e}ta$ and $\hat{\varphi}$ one for the inference and the second one for the correction factor using the variance. This leads to a time complexity of $O(\xi \times 2 \times K \times NN0)$ where ξ is the extra cost for the exponential correction factor. The brute space complexity is around $O(K \times 2 \times (V + D) + NN0)$

In SVM, we carry M documents of size $1 \times K$ for each class. Let's call M the documents/topics pairs during the training stage. The time complexity of SVM is $O(M^3)$ while the memory complexity is $O(M^2)$ where $M \leq D$ Though due to the flexibility of GD and BL in pruning out irrelavant topics, we usually obtain $K' \leq K$, $K'' \leq K$ and $V' \leq V$, $V'' \leq V$under BL and GD. Therefore the memory and time complexities are improved. For instance in CVB-GD-BL(*topic Model I*), we have its memnory complexity as $O(\xi \times 2 \times (K' \times NN0)$ $O(K' \times 2 \times (V' + D) + NN0)$. We could obtain the memory

Table 4.1: Models time complexities

Models	Time complexity
L	$O(\xi \times 2 \times K \times NN0)$
S	$O(M^3)$
L+S	$O(\xi \times 2 \times K \times NN0) + O(M^3)$
I	$O(\xi \times 2 \times (K'') \times NN0)$
II	$O(\xi \times 2 \times (K'') \times NN0)$
I + S	$O(\xi \times 2 \times (K') \times NN0) + O(M^3)$
II + S	$O(\xi \times 2 \times (K'') \times NN0) + O(M^3)$

Table 4.2: Models memory complexities

Models	Memory complexity
L	$O(K \times 2 \times (V + D) + NN0)$
S	$O(M^2)$
L+S	$O(K \times 2 \times (V + D) + NN0) + O(M^3)$
I	$O(K' \times 2 \times (V' + D) + NN0)$
II	$O(K'' \times 2 \times (V'' + D) + NN0)$
I + S	$O(K' \times 2 \times (V' + D) + NN0) + O(M^2)$
II + S	$O(K'' \times 2 \times (V'' + D) + NN0) + O(M^2)$

and time complexities of topic *topic Model II* just by using K" and V'' which the reduced versions of the vocabulary and topics.
LDA does not have ability for to retain the most relevant topics due to its Dirichlet prior. It leads to $O(\xi \times 2 \times K \times NN0)$ $O(K \times 2 \times (V + D) + NN0)$ for LDA as shown in [153]. Here are the tables (above) that recapitulate time and memory complexities of the proposed approach compared to the standard LDA. In Tables 4.1 and 4.2, L=LDA, S=SVM, I= *topic Model I*, and II=*topic Model II*.

We can see that under the time and memory complexities, the LDA is slower and uses a lot of memory than our proposed models. we can also observe that the proposed techniques perform almost equally with their reduced number of topics and vocabulary.

In the worst case, our GD-BL and BL-GD topic models will have the same time complexity as LDA. However, those are very flexible topic models that execute many tasks at the same time including semantic analysis between word and between topics. This suggests they execute each task faster than LDA. LDA does not perform in topic correlation. So it is slower than our proposed models [80]. Tables 4.1 and 4.2 show how the topic correlation analysis improve the time and memory complexities.

4.4 Experimental results

We show the robustness of our proposed approach by selecting some real world and challenging applications in image and text classification. Our framework provides the generative topic models which are then used in SVM. The SVM operates on a series of kernels (in distribution space) such as Bhattacharyya kernel (BK), symmetric Kullback-Leibler divergence kernel (KLDK), the Renyi kernel (RK), the Jensen-Shannon kernel

(JSK), and the Expected likelihood kernel (ELK). In our SVM implementation in this chapter, as we are dealing with a multiclass problem, we select the one-versus-all technique for the training set modeling: that is, the class with the largest positive score will ultimately win the class label. In addition, an 8 fold-cross validation scheme has been implemented to account for the estimation of the design parameters within the SVM.

Using the collapsed Gibbs sampling method and the empirical likelihood scheme, each document distribution is evaluated over the finite set of topics. As we are demonstrating the performance of our proposed approach compared to previous topic models illustrated in Table 4.5 using probabilistic kernels, we also include cases where we compare our proposed topics models to SVMs operating in the original feature space using standard kernels such as linear and RBF. Consequently, we include the performance of our proposed topic models with a linear kernel-based SVM for the text document dataset.

4.4.1 Implementation

This implementation concerns the generative stage where we construct the topic distributions to be utilized by the SVM. Here, we are using the collapsed Gibbs sampling method in variational Bayes inference. The variational update equation is similar to the update equation in the standard CGS. The difference is that here we sample from the variational distribution instead of sampling from the true posterior distribution. Immediately, to deal withn the digamma functions, we can reset the variational update equation using [4] work.

The main idea is to compute the variational model parameters θ_j^c and φ_k using the CVB algorithm which implements this variant of CGS. To do this, we set a number of iterations such that at each iteration we sample a topic for each of the N words in the corpus. We use the variational expected count variables (the variational statistics). We use these statistics to estimate the topic model parameters at the generative stage. The framework requires an intitial setting the variational expected count variables along with the model hyperparameters. We usually set them randomly. Though, many times for the BL hyperparameters, we could also provide initializations in this way: within a class, at the document level we choose $\alpha_{jk} = \frac{1}{k}$ where k $\in \{1, 2, ..., K\}$. We also set α_j such that $\alpha_j \leq \sum_{k=1}^{K} \alpha_{jk}$ or $\alpha_j \geq \sum_{k=1}^{K} \alpha_{jk}$. Then we choose β_j within the same scale as α_j. At the corpus level for BL, we repeat the same process by setting values for λ_{kv} with $v \in \{1, 2, ..., V\}$ and λ and η For the GD hyperparameters at the document level $\alpha_{jk} = \frac{t}{i}$ with $i \in \{1, 2, ..., K\}$ and $\beta_{ji} = \frac{1}{K+i}$ with $i \in \{1, 2, ..., K\}$. At the corpus level we also repeat the same process with λ_{iv} and η_{iv} with $v \in \{1, 2, ..., V\}$. We initialize the number of topics along with the maximum number of iterations. We also randomly initialize the topic assignment associated to each word in the class in the latent $\boldsymbol{z}$ (N-dimensional random variable) associated to each document j. The main expected counts in the sampling process include $\mathbb{E}_q[N_{jk}]$ the number of words assigned to topic k in document j, $\mathbb{E}_q[N_{j(K+1)}]$ the total number of words in topic $K+1$ in document j, $\mathbb{E}_q[N_{kv}]$ the number of times the vth word in the vocabulary is assigned to topic k, $\mathbb{E}_q[N_{k(V+1)}]$ the number of times the $(V+1)th$ word in the vocabulary is assigned to topic k, $\mathbb{E}_q[N_k]$ the the number of times any word is assigned to a topic k $\mathbb{E}_q[N_{j(K+1)}]$ the total number of words in topic $K+1$ in document j. In the document which is a collection of vocabulary words w organized as count data, we associate each word to its initial count (frequency count). In CVB-based CGS algorithm, as shown in Eqs. 135 and 146, we remove the current topic assignment from these update equations by decreasing the count associated to the current assignment. We compute the probability of each topic assignment using Eqs.

135 and 146 leading to a discrete distribution, a K-dimensional variational distribution associated to every word. We sample from this distribution of latent topic assignments and choose a topic that is returned to vector z where it updates the counts. In other words, the appropriate counts are increased. At the covergence, we collect the latent variables z, the variational statistics which allow us to compute the predictive distributions for the document paramter (document-topic), θ_j^c and corpus (topic-work) parameter φ_k.

4.4.2 Text document classification

4.4.2.1 Preprocessing and methodology

In this chapter, we choose a challenging text classification problem using our proposed hybrid technique. For this work, we selected the Yahoo! Answers topic classification dataset. This dataset has been constructed from the original Yahoo! Answers corpus which is a vast collection of text documents containing around $4,483,032$ questions and their corresponding answers (in .csv format). This current dataset has been used in a text classification problem by Zhang et al. in [154]. In fact, the Yahoo! Answers topic classification dataset has 10 main categories (Table 4.3) where the total training set is about $1,400,000$ samples ($140,000$ samples per category). The testing set contains $60,000$ samples (6000 samples per category). The dataset has a 4 column text layout where the first column carries the class labels of each text document. The second and third columns provide questions while the last column shows the best answers to those questions. In our case, in this particular text classification problem, we are interested in documents containing answers and their corresponding classes from the corpus layout.

Though, in this experiment, we did not use the whole dataset as we utilized only a subset of the data that consists of 6000 samples per category, so 60000 samples in total. This is mainly due to poor initializations which were slowing down the sampling process. We reduced the size to speed up the sampling process.

As usual for text document data, the collections are initially unstructured or noisy as they carry a lot of unwanted materials. Consequently, in the preprocessing stage, we cleaned up the data by removing irrelevant items such as stop words and punctuations through MATLAB. In each class, 90% of the dataset have been assigned to the training set while the remaining is the testing set. The training set obtained is then used to construct the bag of words from the tokenized documents. Further preprocessing steps have been implemented to remove infrequent words in BoW model (for instance, words that appear less than two times in documents). In addition, empty documents have been also removed from the training data. The characteristics of the training set following the BoW framework is summarized in Table 4.4 which shows the frequency count data represented in a matrix form, the total number of documents, the vocabulary size, and the total number of words in the corpus.

The frequency count data (training set) is then used by our algorithm where we learn documents topics first: this is the generative stage. It is important to mention that our text data using the BoW framework is really sparse due to the large size of the vocabulary. We proceeded with a sparse-based data representation for efficient storage management in this batch processing.

For the generative framework, we finally obtain the optimal number of topics at $K = 60$. Once our generative topic model is built, we represent each document as a topic distribution.

We, in fact, constructed two generative topic models: the *topic model I* or CVB-GD-BL-based topic model and *topic model II* or CVB-BL-GD-based topic model, all presented in section 4.3. With these topic models, we estimated the predictive distributions that allow us to express the document distributions using Eq. 139. The topic distribution are then used by the SVM classifier to perform document categorization with probabilistic kernels. The representation of each document as probability distribution is fully detailed in subsection 4.3.3 in this chapter. Importantly, there are no clustering and quantization steps for text documents during the BoW formation. These steps only occur when dealing with images in feature representation. Text documents naturally decompose into bag of words. This ultimately summarizes our generative-discriminative approach for text document classification.

4.4.2.2 Results

Initially, the BoW representation of the data shows a very sparse data, and the first samples used for modeling did not yield good approximates. It means there is a need to provide more discriminative features that facilitate classification. As we increase the size of the documents and the number of latent factors or topics as shown in Figs. 4.2, 4.3, and 4.4, we saw an immediate improvement in the estimates for the topics and the distribution over the topics. The improvement in the estimates not only shows the characteristics of each document, but also exhibits differences between these documents by observing the topic and distribution structures.

The experimental results from our proposed approach with this text dataset using the different probabilistic kernels utilized in this chapter are summarized in Table 4.5. These results show that our two generative (topic) models implemented, (*topic model I* and *topic model II*) have provided satisfactory performance with SVM framework compared to their major competitors (such as LDA, CVB-LDA, CVB-LGDA, and LGDA) in this text document classification. So, the hybrids obtained from the proposed topic models outperformed their competitors under these probabilistic kernels.

All the hybrids in this text document classification have successfully provided good results with the expected likelihood kernel (ELK) which is a linear probabilistic kernel. Under the ELK-based SVM, the *topic model II* had the highest accuracy (68.53%). In overall, results from hybrids using *topic model II* were slightly improved compared to hybrids from *topic model I*. In this experiment, the linear kernel was able to outperform nonlinear kernels in text document classification. We think that linear probabilistic kernels could be seen as alternatives to nonlinear probabilistic kernels in text document classification. Though, the JSK-based SVM coupled with *topic model II* remains the best performer among nonlinear probabilistic kernel models.

This ultimately demonstrates the robustness of probabilistic linear kernels in text document classification. However, both topic models proposed in our work outpformed an SVM-based classifier using traditional and standard linear kernel (in the original feature space). The classification accuracy with topic Model I is 58.43%. It is 56.38% with topic Model II, and 54.41% with SVM.

Finally, these performances ultimately illustrate the importance of documents representation in distribution space. And this starts from providing an optimal number of topics from our generative models from which distributions are built over the topics structure. The ability to summarize documents (initially represented in 8805 dimensional feature space in this chapter due to the size of the vocabulary within the BoW as shown in

Table 4.4) using efficient and very few low dimensional features such as topics is an ideal framework for memory space management in databases. From documents with initially 8805 features, we obtain at the end $K = 70$ topics from the generative model to represent documents new features in distribution space.

Table 4.3: 10 Categories for text documents

Text Document Categories	Class label
Society & Culture	1
Science & Mathematics	2
Health	3
Education & Reference	4
Computers & Internet	5
Sports	6
Business & Finance	7
Entertainment & Music	8
Family & Relationship	9
Politics & Government	10

4.4.3 Natural scene categorization dataset

4.4.3.1 Preprocessing

In this experiment, we are performing image classification using our proposed hybrid framework. We also used the well-known natural scenes dataset that has 15 categories as shown in [48]. It is a challenging dataset. Here is the list of the classes along with their size: (Suburb, 241), (Living room, 289), (Coast, 360), (Forest, 328), (Highway, 260), (Mountain, 374), (Street, 292), (Office, 215), (Store, 315), (Bedroom, 216), (Inside city, 308), (Tall buidling, 356), (Open country, 410), (Kitchen, 210), and (Industrial, 311). The corpus as illustrated in Fig. 4.5 has in total 4485 images. The dataset is also a collection that contains different categories (for instance, mountain and highway) as well as similar categories (for instance, the 4 indoor categories such as office, living room , kitchen, and bedroom from [2]) to fully characterize the concept of interclass and intraclass variation problems.

From each category, the dataset is split in two groups: the testing set carries 100 samples while the training set gets the remaining. This is similar to our previous work with this data in [80, 67] where we used the BoW method to transform the SIFT (scale invariant feature transform) descriptors (from image patches) into codebook or vocabulary after clustering

Table 4.4: BoW information for the text document modeling

BoW	Characteristics
Total Counts	53087 × 8805
Vocabulary	[1 × 8805 string]
Total Number of Words	8805
Total Number of Documents	53087

Table 4.5: Hybrid models performances for the text document dataset

%	BK	KLDK	RK	JSK	ELK
topic model I	61.45	62.16	62.27	63.49	67.51
LDA	45.56	48.67	49.25	50.67	57.89
CVB-LDA	46.12	49.87	57.43	54.89	55.57
CVB-LGDA	50.78	51.65	52.10	53.09	57.16
LGDA	48.36	48.98	49.67	50.18	56.54
topic model II	63.32	63.67	65.74	66.19	68.53

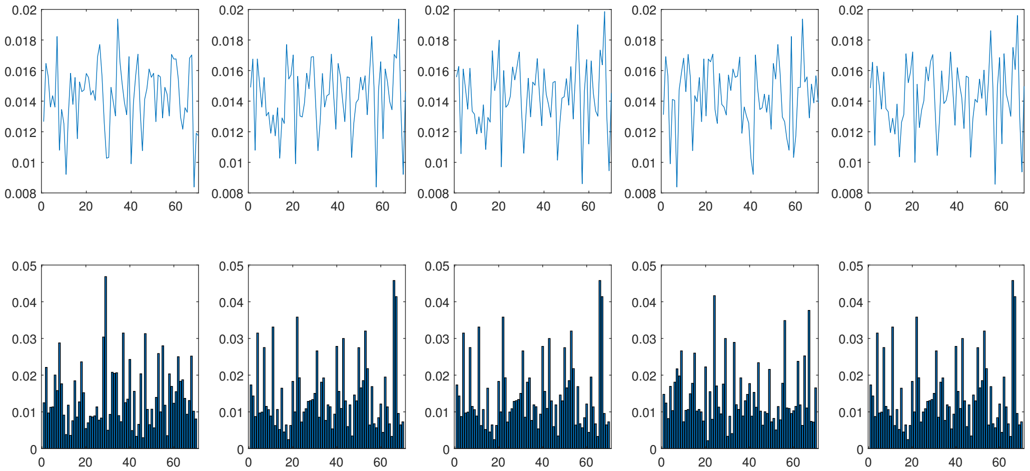


Figure 4.2: Processing results from increasing documents size and the number of topics

and quantization process [2, 11, 80, 67]. The training set (count data) obtained is then used to build our generative topic models with asymmetric priors. Following the steps in the text classification problem in subsection 4.4.2, we characterize each document distribution using subsection 4.3.3. These documents are then used by our SVM which performs with probabilistic kernels. It is noteworthy that based on our previous work [80, 67], the optimal number of topics and vocabulary size are reached at $K = 90$ and $V = 1000$, respectively for the implementation of our generative topic models. This is because of the ability of the GD and BL in pruning irrelevant topics and vocabulary size. We therefore obtained a model selection with very reduced number of topics and vocabulary size.

4.4.3.2 Results

We showed earlier the low performance of the hybrids with the expected likelihood kernel (ELK): 58.43% for topic Model I, 56.38% with topic Model II. This probabilistic linear kernel was not able to carry enough discriminative information or features that could enhance performances in this image categorization problem. In general, from our experiment, nonlinear probabilistic kernels used in this hybrid generative-discriminative setting have been observed to outperform the ELK. The two topic models in our proposed approach combined with nonlinear probabilistic kernels-based SVM show robustness of our methods

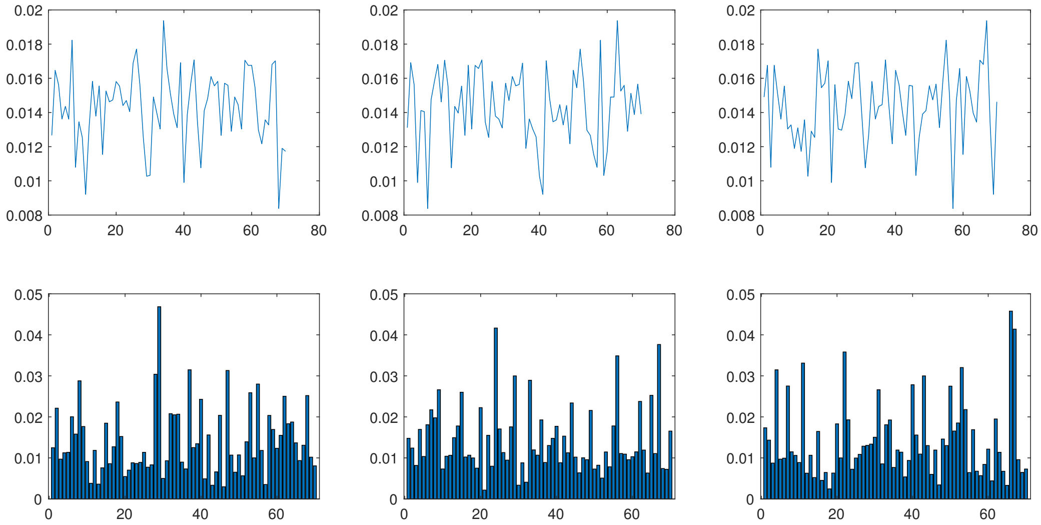


Figure 4.3: Three classes from text corpus documents with associated topic structure

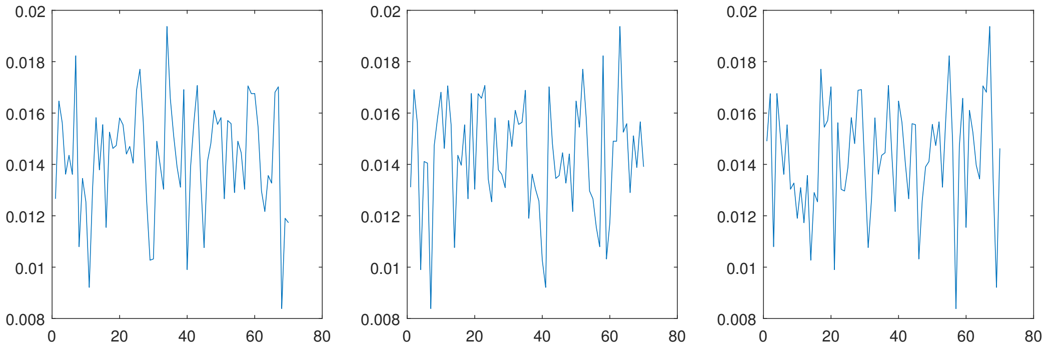


Figure 4.4: Multinomial distributions from 3 text documents of different classes

with a result around 85% in accuracy from *topic model II*. These two hybrids in our scheme seem to equally perform well with nonlinear probabilistic kernels especially the JSK. They both outperform their competitors such as LDA, CVB-LDA, CVB-LGDA, and LGDA. These results show that nonlinear probabilistic kernels are robust and efficient in image classification than in text categorization. In this experiment, nonlinear kernels are able to characterize the intrinsic properties in images than linear probabilistic kernels represented by the ELK. This justifies the poor performance in the ELK for its inability to adapt to changes in view and illumination in images for instance since such phenomena induce nonlinearity in the dataset resulting in changes in document distributions. This instability in the distributions has a negative impact on linear probabilistic kernel function (ELK).

In addition, the proposed topic models (implemented in this chapter) performances have been compared to a Gaussian or RBF kernel-based SVM classifier which operates in the original feature space (75.35% with *topic Model I* and 76.65% with *topic Model II*). The SVM with RBF kernel using orginal feature instead topic distribution provided an accuracy of 68.27%. These topic models outperform the RBF-based classifier. The performance of

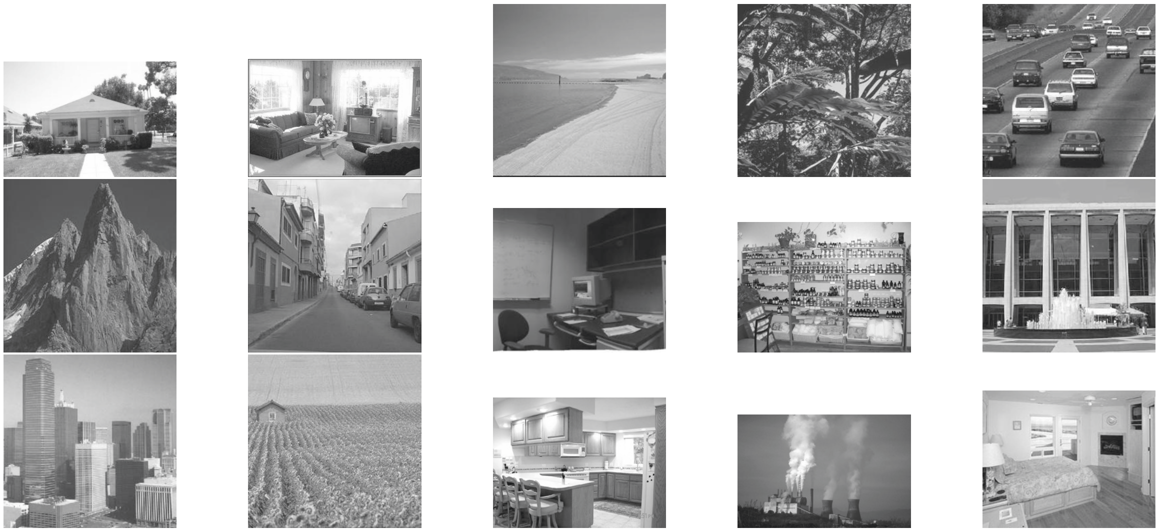


Figure 4.5: Examples from the natural scenes image dataset (15 categories).

our method could also be explained by the robustness in the generative topic models for their ability to characterize effectively the documents as probability distributions with a better parameterization. For instance, a random selection of 5 documents has been made whitin the natural scene category dataset. As shown in Figs. 4.6 and 4.7, and similar to the scenario presented in our text document classification, the first row, in each figure, illustrates the convergence process while the second row exhibits the word distribution in the documents. The last row provides the topic structure in each document. Under our proposed approach, we can see that the documents are different according to their classes. In Fig. 4.6 for instance, on the second row, documents 1, 2, and 4 have similar topics and similar distributions over topics. Still on the second row, same observations could be made about documents 3 and 5. These 5 documents ultimately belong to 2 classes from their distribution characteristics. This robustness in approximating effectively the generative topic model facilitates the task for the probabilistic kernel to perform accurately as it measures similarity between distributions within the discriminative framework. As we start increasing the size of the dataset, the number of topics, and the size of the vocabulary during training, we notice improvement in the results with our proposed hybrids. The final optimal number of topics and size of the vocabulary are obtained at $K = 90$ and $V = 600$, respectively. This constitutes the characteristics of the generative approach that we use to construct our discriminative classifier.

Table 4.6: Hybrid models performances for the natural scenes dataset

%	BK	KLDK	RK	JSK	ELK
topic Model I	78.31	79.18	82.17	82.32	70.65
LDA	59.34	65.54	68.67	69.43	55.41
CVB-LDA	65.38	70.3	70.86	71.57	57.85
CVB-LGDA	70.51	69.96	75.71	80.53	68.34
LGDA	68.45	70.43	75.35	77.64	63.50
topic Model II	78.54	78.98	80.78	85.47	74.67

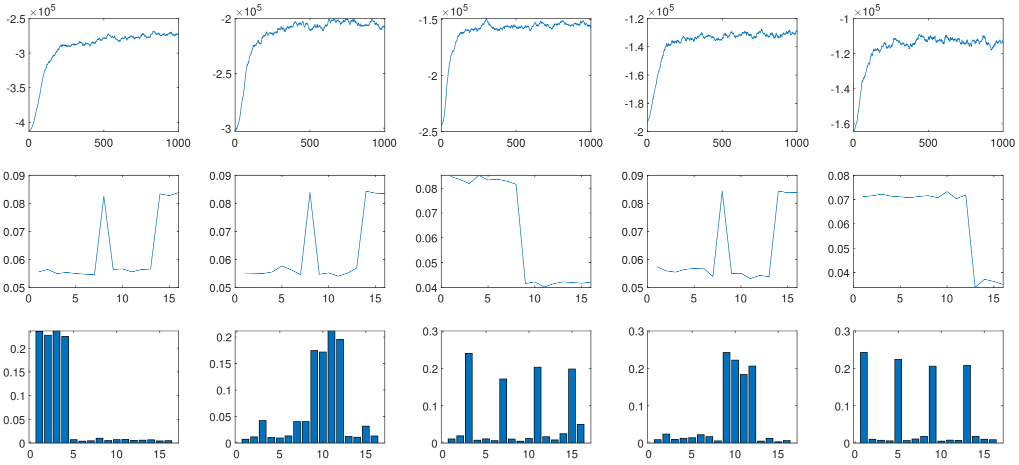


Figure 4.6: Five image documents in natural category scene dataset

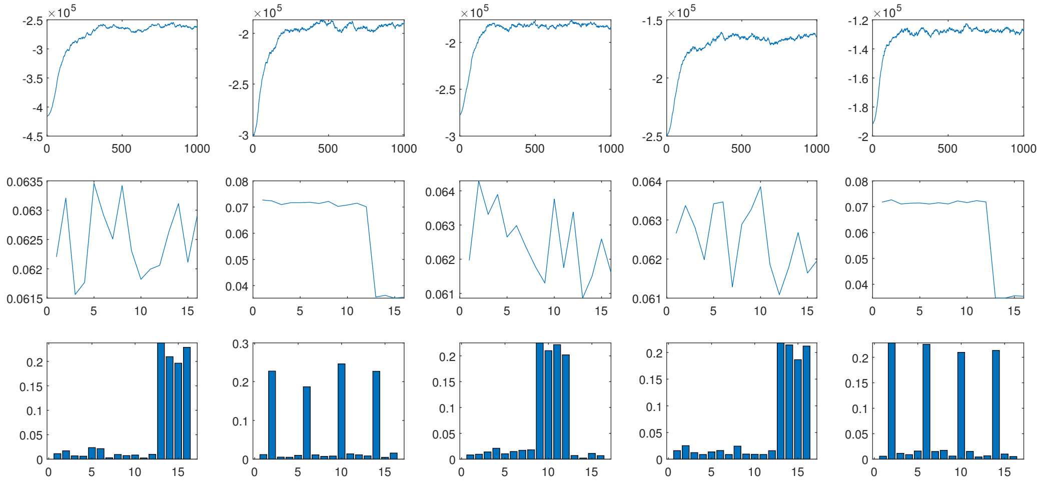


Figure 4.7: Analysis of 5 image documents in natural scene category dataset

4.4.4 COREL dataset

For this second experiment of image classification using our proposed method, we selected the COREL database as illutrated in Fig. 4.8 from the Corel Photo Gallery [155] for our image classification framework. Over thousands images, the collection contains animals, airplanes, cars, plants, landscape and textures, artistic objects, vehicles, and people. The database has in fact been summarized into 80 categories in total containing 8000 images (100 images per class). Each image in the collection has approximately a size of 325 × 255, in JPEG format.

Initially, for feature extraction method, we decided to follow the method implemented in [110] to collect the low frequency features provided by the DCT (Discrete Cosine Transform) from the patches obtained by the sliding window process over the images using MATLAB.

These low frequencies in DCT are specialized in capturing relevant characteristics in images. As the generative topic model in our implementation was struggling to be successful with this feature extraction scheme, we decided to use SIFT features similar to the work in [2, 11, 80, 67] and the one in the previous section in this chapter about natural scene categorization. In this work, we used all the 80 categories. The SIFT method and BoW architecture are described in [2, 11, 80, 67] for image representation in feature space.

Once the generative topic models are implemented, we use probabilistic kernels to carry the document topic features to the SVM for classification. The technique ultimately requires the representation of each document in the distribution space to facilitate the work for the probabilistic kernel machine. Then afterwards, we compare the performance of our proposed approach to its competitors in topic modeling. We also maintain an optimal number of topics at $K = 70$ for a vocabulary size of $V = 600$ for the implementation of the generative topic models.

The implementation of our method has shown the performance of the hybrids with nonlinear probabilistic kernels compared to linear probabilistic kernels such as ELK (expected likelihood kernel). From the results (in terms of accuracy) obtained, we can observe that these hybrids performances with ELK were less improved compared to the case of nonlinear kernels such as BK, KLDK, RK, and JSK. This is translated into a low accuracy value for the ELK. The hybrids provided by our proposed generative approach, (*topic model I* and *topic model II*), with the probabilistic kernel-based SVM have demonstrated higher results. The combination *topic model I* and SVM showed an accuracy of 79.83% with the JSK. In overall, these two topic models perform equally within the discriminative setting especially with the JSK.

Similar to the natural scene document modeling case in the previous section, in this COREL dataset also, we randomly selected 5 documents (Figs. 4.9 and 4.10). Our proposed topic models were able to show the efficiency of the representation of documents as distributions. Through these distributions characteristics, the documents were able to exhibit their differences. Here, each of these documents (by observing the second row) belongs to a different class as illustrated in Figs. 4.9 and 4.10. Our generative models implemented have shown better performance when compared to an RBF-based SVM classifier in the original feature space (*topic model I* with an accuracy of 72.70% and *topic model II* with 70.40%). Implementing the SVM in the original space provided 65.34% as classification accuracy. By using topics, we were able to provide a lower dimemsional space that allows a better compression of the data. The low dimensional space is used to represent the documents.

Table 4.7: Hybrid models performances for COREL dataset

%	BK	KLDK	RK	JSK	ELK
topic Model I	75.51	76.39	77.98	79.83	67.42
LDA	57.65	60.56	67.43	68.36	55.39
CVB-LDA	60.45	63.78	68.56	69.43	57.54
CVB-LGDA	63.42	65.45	70.12	71.48	58.29
LGDA	62.10	64.33	68.27	70.25	58.87
topic Model II	74.10	74.87	77.21	78.75	70.38

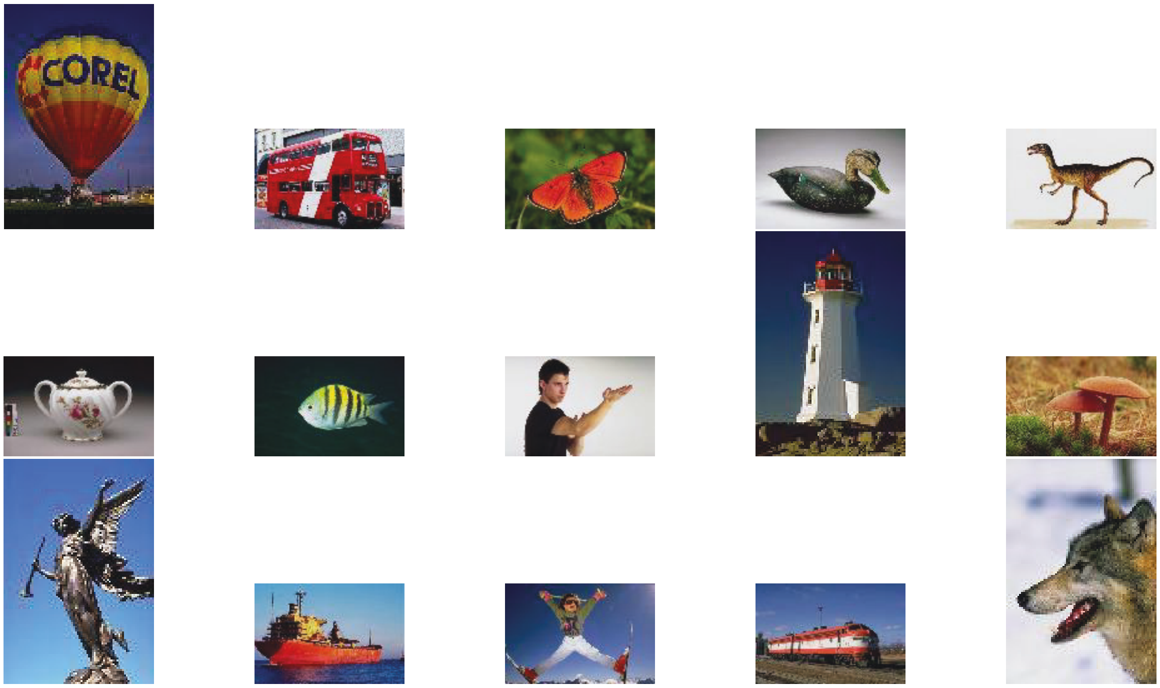


Figure 4.8: Corel dataset (15 out of 80 categories)

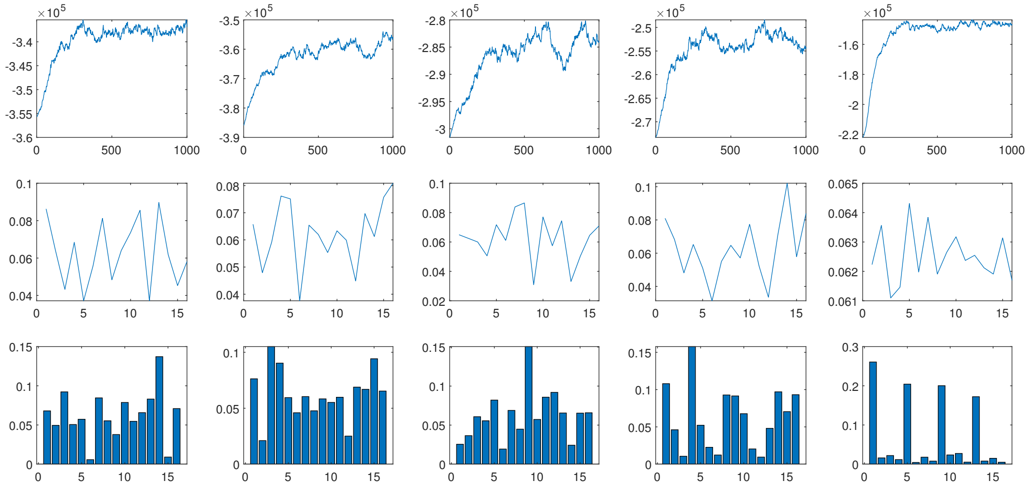


Figure 4.9: Analysis of image documents in Corel dataset

4.5 Conclusion

In this chapter, we demonstrated the effectiveness of documents or data representation (generative features) from the proposed topic generative framework coupled with the implementation of powerful probabilistic kernels-based SVM classifiers that provided good performance in classification. The use of asymmetric GD and BL conjugate priors simultaneously (within the same generative process) in our topic modeling framework

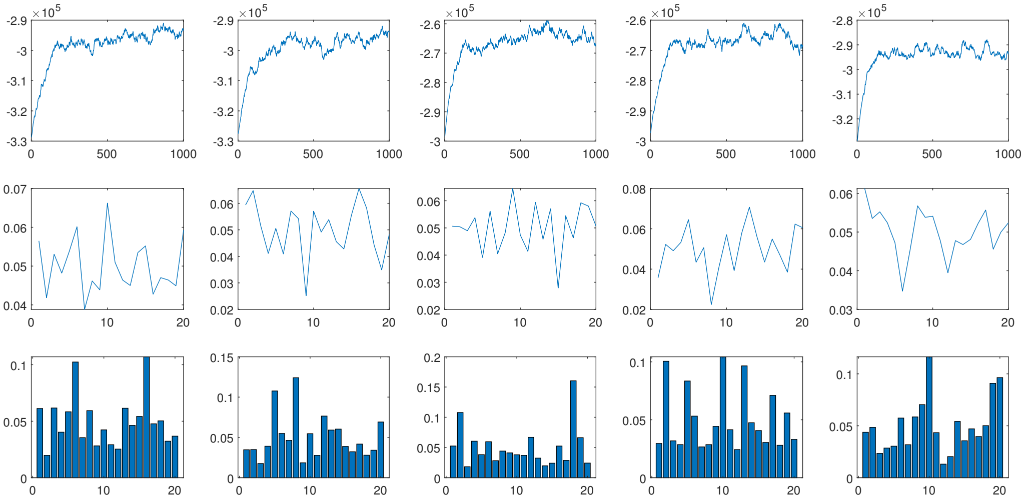


Figure 4.10: Characteristics of image documents in Corel dataset

led to two models: the CVB-GD-BL-based topic model (*topic model I*) and the CVB-BL-GD-based topic model (*topic model II*). This ultimately characterizes the generative-discriminative setting in our proposed approach. The discretization of the continuous visual feature space due to clustering and quantization schemes for the formation of the visual codebook led to the reformulation of probabilistic kernels from continuous space to discrete space as we deal with empirical (discrete) distributions. Using some challenging datasets in machine learning and computer vision, we are able to extract intrinsic characteristics from text and image documents for the implementation of our hybrid models. Topic representation is an effective summarization method to allow topic models to work in finite dimensional spaces (low dimensional spaces). This automatically presents the advantage of solving memory space (storage) issues in databases. In other words, the space complexity is refined and improved within our proposed framework.

The implementation of generative models in the fully collapsed space of latent variables provided a framework (sampling) that allows the computation of probabilistic kernels through empirical likelihood scheme. This setting facilitates the representation and parameterization of our documents (texts and images) as distributions for the kernel machine. This representation has been beneficial for the modeling of our hybrids as documents now have ability to carry effectively local information from generative topic models into discriminative classifiers that operate with distributions. Distributions are always seen as accurate and compact representations of the data since they can efficiently hold some useful properties such as semantics within the observed data. This reality is demonstrated in our experiment as we successfully show that despite the performance of standard kernels-based SVMs in the original feature space, probabilistic kernels-based SVMs provide the best performance and results especially when combined with robust topic models. These characteristics illustrate the effectiveness of our hybrid models and their performance within a wide variety of datasets showing therefore the ability for our proposed framework to generalize. The fully collapsed representation was also key to the success of our generative approach by connecting a hybrid inference (the collapsed variational Bayes, seen as one of the state-of-the-art inference techniques in topic modeling with its flexibility to combine both the performance of the variational Bayes and the collapsed Gibbs sampler) to hybrid model (generative-discriminative). The hybrid techniques using CVB-LGDA and CVB-LDA in this generative-discriminative approach have shown better performances compared to the LDA-based hybrids in uncollapsed space.

As generalized Dirichlet and Beta-Liouville distributions are more flexible than the Dirichlet, using these priors in topic modeling presents some advantages in the generative-discriminative setting. This ultimately justifies the good performance in our proposed approach as we implement our topic models with these two different priors (asymmetric) used simultaneously within the same generative process. Compared to previous hybrid models, our proposed approaches mostly outperform them in our datasets. As a result, the edge is given to our current proposed methods. With the right probabilistic kernel, the hybrid methods from topic Models *I* and *II* could also perform almost similarly with the majority of our datasets in a sense that they both provide mostly, robust and coherent generative topic features to the SVM as shown in the performance results compared to their competitors. However, within our proposed methods, the hybrid, *topic model II/SVM* provides a better performance in terms of time complexity in comparison to the hybrid, *topic model I/SVM*. This is mainly due to the intrinsic characteristics of the (asymmetric) Beta-Liouville conjugate prior for the document parameter's modeling besides robustness and flexibility. To its advantage, the distribution (BL) has generally few parameters compared to the GD. As a result, inferences were observed to be faster with the hybrid *topic model II/SVM* as it effectively characterizes or models the document parameter with (asymmetric) BL while also providing robust generative features to the kernel machine. This is in contrast to the hybrid *topic model I/SVM* which samples the document parameter from (asymmetric) GD, and it is observed to be slower in estimations despite its robust performance. The relationship between our topic generative features and kernel formulations for SVM also demonstrate that our nonlinear probabilistic kernels implemented performed well with images than linear probabilistic kernels such as ELK. Images often provide features that are too complex to be linearly separated. Changes in view and illumination for instance could have impacts on image feature characteristics and therefore on the distributions. Nonlinear probabilistic kernels have ability to adapt to these changes better than linear kernels. On the other hand, text documents classification tends to be well characterized with linear probabilistic kernels. Our models were able to exhibit these characteristics through our datasets showing therefore the robustness of the framework. This explains the importance of knowledge about the data as it can influence the choice of the kernel functions in the discriminative framework. Therefore, the strong performance of the JSK (Jensen-Shannon kernel) on our proposed topic models could be explained by the capability of this nonlinear probabilistic kernel in handling and characterizing effectively generative features represented as empirical distributions such as the ones implemented in our topic models.

We witnessed, during implementation that the models require many parameters and hyperparameters to be initialized. The complexity of the models has been increased, and initializations affect the results. Importantly, our proposed approach remains an alternative to nonparametric models in finite dimensional space (with finite mixtures) for classification. However, as topic models in finite dimensional space always struggle in providing very efficient and accurate model selection criteria, a future work could be about investigating on the possibility to implement a nonparametric model due the high complexity in our datasets. We could also emphasize on inference based on hyperparameter estimation to reduce problems related to poor initializations.

Chapter 5

A Two-Level Hierarchical Latent Beta-Liouville Allocation for Large Scale Data and Parameters Streaming

As an extension to the standard symmetric LDA (latent Dirichlet allocation) topic model, we implement asymmetric Beta-Liouville (BL) as a conjugate prior to the multinomial and therefore propose the MAP (maximum a posteriori) for LBLA (latent BL allocation) as an alternative to maximum likelihood estimator (MLE) for models such as PLSI (probabilistic latent semantic indexing), unigrams, and mixture of unigrams. Since most Bayesian posteriors, for complex models, are intractable in general, we propose a point estimate (the mode) that offers a much tractable solution. The MAP hypotheses using point estimates are much easier than full Bayesian analysis that integrates over the entire parameter space. We show that the proposed MAP reduces the three-level hierarchical LBLA to two-level topic mixture as we marginalize out the latent variables. In each document, the MAP provides a soft assignment and constructs dense EM probabilities over each word (responsabilities) for accurate estimates. For simplicity, we present a stochastic at word-level online EM (expectation-maximization) algorithm as an optimization method for MAP-LBLA estimation whose unnormalized reparameterization is equivalent to a stochastic CVB (collapsed variational Bayes). This implicit connection between the collapsed space and EM-based MAP-LBLA shows its flexibility and helps in providing alternative to model selection. We characterize efficiency in the proposed approach for its ability to simultaneously stream both large scale data and parameters seamlessly. The performance of the model using predictive perplexities as evaluation method shows the robustness of the proposed technique with text document datasets.

5.1 Introduction

In topic modeling literature, the classical maximum likelihood (ML) estimator has been applied to several classic topic models including PLSA (probabilistic latent semantic analysis), unigrams, and mixture of unigrams. Because of its frequentist nature, it is very limited in predictive modeling as it does not consider prior information. Reduced to multinomial distributions with no prior information, these classic topic models fundamentally carry the limitations of multinomials: using only frequencies as ways to represent probabilities often leads to poor estimates. In a highly sparse dataset, without any smoothing, frequencies are more likely to assign zero probability for unseen or rare events. Moreover, and very often, multinomials do not capture efficiently the words burstiness because of the lack of priors [6], [36, 7]. The integration of prior information has become fundamental for the flexibility of topic models such as LDA over classical frequentist approaches. In other words, the limitations of classic frequentist models led to the emergence of the LDA and its variants. LDA is a latent generative probabilistic graphical model that assumes that words are generated from a mixture of topics [8]. The topics are themselves distributions over the vocabulary words. The topic proportions vary from one document to the other and exhibit how documents are organized, summarized according to the global topics. LDA allows documents to exhibit multiple topics.

The success of the LDA model has reinforced its use in a wide variety of applications mainly in text document analysis [92, 156, 135] and computer vision [101, 103, 97]. Compared to LDA, in a unigram model, words in a document are drawn from a single multinomial distribution (the word simplex). The mixture of unigrams is an augmented version of the unigram model with a discrete topic (latent) variable. With the mixture of unigrams, a document is now generated from only a single topic [5]. The PLSA is almost similar to LDA topic model but with no prior information [4]. It is a relaxation of the mixture of unigrams assumption as it allows a document to exhibit multiple topics. As presented earlier, the lack of priors in PLSA makes the model unfit for prediction and often suffers from overfitting problems. The LDA topic model provides a solution to the PLSI or PLSA, unigram, and mixture of unigrams by including prior information as it treats topic proportions as random variables.

Since LDA and its current variants rely extensively on prior information, it is natural to perform parameter estimation where the logarithm of the priors offers a possibility to act as a regularizer of ML estimates. This ultimately introduces the flexibility of the MAP framework. The MAP estimates are point estimates whereas full Bayesian analysis often characterizes the posterior mean instead of a single estimate (mode) [157]. However, point estimates are often preferred because posterior means require computationally expensive methods and often lead to intractable solutions. The MAP framework models directly the posterior distribution of the parameters.

Due to its prior information, the MAP is robust to outliers. In topic modeling, these advantages could present a possible MAP technique with standard EM algorithm in online fashion as alternatives to complex methods such as variational Bayes (VB)[3, 5], MCMC (Markov Chain Monte Carlo) using CGS (collapsed Gibbs sampling) [158], and EP (expectation propagation)[159]. Although, the MAP is not invariant to reparameterization as it requires the Jacobian information to relocate the mode, we can use unnormalized reparameterization to simply seek for equivalent models that could help in characterizing efficiently its online framework [13, 153, 4]. We can also observe from the literature that online LDA topic models such as stochastic CVB (SCVB) [13], online CGS (OGS) [160], [63], SVB (stochastic variational Bayes) [1, 65], online VB (OVB) [65], generally implement stochastic optimizations [161] from batch LDA models (VB[3],[5], CVB [12] and CGS [162, 162, 157, 163]). Furthermore, as these models only focus on large scale data processing while ignoring parameter streaming, the work in [17] recently implemented an online LDA topic called fast online EM that accommodates parameter streaming to large scale data modeling. Modeling dependency between hidden variables has been ignored in standard variational Bayes that assumes that joint variational distributions variables are all independent from each other. In addition, relaxing the mean-field assumption can become extremely challenging in the latent update equations when using empirical Bayes framework (the log marginal distribution) to set the lower bound as shown in [12] with its symmetric LDA.

The first critics to all these methods always point to the use of Dirichlet (Dir) prior in LDA for inference. LDA assumes that its topic components are all independent when using the Dir prior. Furthermore, one of the limitations within the multinomial assumption is that often the poor estimates are results of the fact that events are supposed independent, which is not always the case [164, 165]. By choosing flexible priors instead, we could characterize efficiently dependency between events which are translated into dependency between documents and topic components (topic proportions). LDA is not the right model when it comes to characterizing dependency since it systematically prohibits such interpretation because documents simply cannot be dependent under the LDA topic model. As suggested in [8] using Dir leads to an unrealistic way to explore unstructured collections of documents because in real life scenario, there is always a high probability of existence of a topic correlation setting in a large collection. This drawback in LDA promoted the use of logistic normal distribution as an alternative to the Dirichlet prior in topic correlation [121, 120, 119, 125, 101, 21, 22, 113, 114]. Another major problem and setback in finite mixture topic modeling is the lack of effective model selection criteria [5, 9, 17, 41, 80, 8] especially with LDA which relies on cross-validation solutions. For large scale applications cross-validation methods are not efficient. Since LDA is too restrictive due to the Dir distribution while non conjugate priors such as logistic normal distributions often led to very complex deterministic (VB, CVB, and EP) and MCMC using CGS inferences, we propose a very simple algorithm that performs a MAP estimate on the LBLA where the conjugate prior to the multinomial is the asymmetric Beta-Liouville (BL) prior. The flexibility of the prior allows us to model dependency between documents. In attempt to induce dependence, the CVB marginalizes out the parameters, while leaving the latent variables; on the other hand, the proposed MAP-LBLA integrates out the latent variables instead and even reduces the three-level hierarchical LBLA topic model to just two levels. This ultimately simplifies computation.

We proposed a stochastic at word-level online EM algorithm for MAP-LBLA as an alternative to online LDA in [17] to which we provide a refined model selection including data and parameter streaming for fast inference. Our model outperforms the LDA-based topic models and shows the robustness of the scheme in producing very accurate predictive distributions and perplexities. In our method, because implementing a word-level processing, documents parameters and topics are global parameters. This is in contrast to the standard stochastic VB (SVB) approach that supports a document-level processing where the only global parameter is the topic. We show that our stochastic algorithm using online EM has connections within the collapsed variational Bayesian inferences through unnormalized reparameterization of the MAP [4, 13]. Under this reparameterization it is clear that our technique could follow a minibatch of size one as we will show later in the coming sections. This allows the model to manage the vocabulary size easily. As we implement a stochastic method that favours small samples at a time in a document, the MAP can effectively regularize MLE estimates and performs better than frequentist estimators. To each word accessed, the E-step provides a sample (EM responsibility vector) from the posterior distribution; but no longer stores it within our stochastic method as in the batch EM. All these flexibilities make our approach more robust and accurate over extremely fast methods that could escape many critical steps (during processing) that are required for a good modeling. We finally demonstrated that our model while using unnormalized update equations is flexible due to the asymmetric BL prior that generalizes the Dir distribution. The main contributions of our proposed parametric topic model are:

- We provide alternative to the MAP-LDA and its variants including stochastic and online versions. We selected the BL prior to estimate very heterogeneous topics that enhance predictive models and perplexities.

- The simplicity of inference with the standard EM algorithm over complex methods such as variationals and EP including MCMC methods such as CGS and CVB allows to model dependence in exact manner which leads to much accurate parameters estimates.

- We successfully provide a solution (alternative) to model selection problem within finite mixture topic model setting which is a very challenging concept due to the lack of criteria for model selection in topic modeling in general as our model stochastically favors small samples which are regularized by the prior information within the proposed MAP framework: our approach uses its equivalent models to efficiently propose model selection

This chapter is organized as follows: section 5.2 presents the related work and background while section 5.3 introduces the proposed online EM-based MAP-LBLA approach. Section 5.4 illustrates the experimental results and finally section 5.5 provides future work and conclusion.

5.2 Related work and background

LDA is a generative probabilistic graphical model that summarizes documents (texts, images) as mixtures over topics. Topics are distributions over vocabulary words [3]. Under its generative process, LDA assumes that a word is generated from a mixture of topics [8]. Many inferences support the LDA architecture and make it the most recognized topic model in the literature. The main inferences include VB and CVB which describe the variational approaches while GS (Gibbs sampling) and CGS (collapsed Gibbs sampling) which are MCMC methods [12]. The CVB and CGS are based on collapsed representation where the parameters are marginalized out: CVB is variational method in the collapsed space, therefore a deterministic approach where CGS is an MCMC method or stochastic in the collapsed space. The CGS provides a hard assignment technique while CVB favors a soft clustering method resulting in a K-dimensional variational distribution being associated to each word or token [157]. One of the advantages of the collapsed representation was to characterize a dependence structure in topic modeling as a way of relaxing the independency assumption in mean-field variational methods. It also provides a better lower bound to the log marginal likelihood for accurate predictive distributions showing parameters estimated in exact way [12]. The VB and GS are inferences in uncollapsed spaces. The work in [68] has constructed a partially collapsed space where documents proportions are marginalized out leaving the latent variables and the topics.

The majority of these batch inferences have been extended for online processing leading to OVB [65], SCVB [13], OGS [160, 63], SVB [1, 65], etc. For a direct modeling of the parameters, the MAP marginalizes out the latent variables and optimizes an EM lower bound on the posterior distribution of the parameters in M-step. The E-step follows a stochastic expectation that computes unnormalized expected sufficient statistics (for exponential family distributions) also called EM statistics [13]. In latent topic models, we can observe that the MAP integrates out the latent variables while the CVB inference marginalizes out the parameters. Authors in [4] have tried to show the connection between these inferences for LDA through hyperparameter analysis. MAP reduces the three-level topic model to two levels and introduces a mixture model setting. Other main characteristics and challenges of LDA model include the problem of a robust model selection [5, 9, 17, 41, 80], and correlation between topics [8, 20, 166, 113, 114]. The problem with these inferences is that the majority are LDA-based approaches. Furthermore, LDA could not characterize dependence structure because it is one of its intrinsic limitations. Under the Dirichlet its random variable components are independent, so correlation between topics could not be emphasized with efficiency within the LDA. The model selection framework in finite topic modeling is very challenging. For instance, the multitopic technique [41] is efficient for batch learning but not for online one. Its limitation is due to the relevance feedback from a user. It means it cannot perform without human intervention. The VI (variation information) method [9] operates within the uniform probability measure setting which we believe could not be ideal for the MAP technique because estimate with uniform priors are equivalent to ML estimates.

The fast online LDA topic model in [17] provided a model selection that uses accumulated residuals (to select the number of topics and vocabulary size) combined with a buffering system that facilitates easy transfer of data between the PC (personal computer) memory and its external storage. Its sorting mechanism based on residuals for model selection is complicated because both the time and memory (space) complexities rely on the number of topics K and vocabulary size V. Despite the fact that the updating and normalization steps of the responsibility vector benefit from time complexity, it is really difficult to understand how the framework became invariant to the number of topics at some points when analyzing the time complexity.

Due to these difficulties, we propose an alternative with more improvement: we implement an online EM method for MAP estimation with LBLA topic model, a generalization of the LDA. The proposed approach uses a BL prior [11, 64, 107] as an alternative to the Dir distribution. The BL has ability for topic correlation [107] framework similar to work in [113, 8, 20]. We emphasized on a word-level stochastic online EM approach whose unnormalized parameterization connects with stochastic inferences in the collapsed space. Our proposed method uses its internal structure to reduce the number of topics and vocabulary size and allows for efficient data and parameter streaming. Importantly, compared to other methods that use computationally expensive resources for model selection, our proposed model selection technique does not require too much computation. Its advantage is that it promotes small samples processing (reasonable minibatch sizes) which encourages the use of small number of topics and vocabulary sizes. This reason explains our stochastic method which can implement a minibatch setting of size one for small samples. It constantly processes and updates, for the global topic matrix of size $K \times V$, only its vth column using a reduced number of topics. Small samples are appropriate for our MAP-LBLA method because of the presence of the prior to regularize or correct estimates.

Table 5.1: Variables and definitions

Model variables and acronyms	Definitions
$\mathscr{D}$	Total number of documents
$\mathscr{W}$	Total number of words in the corpus
$\mathscr{V}$	Minibatch size
$\mathscr{W}_j$	Total number of words in a document j
K	Total number of topics
V	Vocabulary size
(i, j)	The ith word or topic assignment in the jth document
k	The kth topic
$X = \{x_{ij}\}$	Observed words
$Z = \{z_{ij}\}$	Latent variables
$\theta_j = \{\theta_{jk}\}$	Topic proportions
$\varphi_k = \{\varphi_{kv}\}$	Corpus parameters or global topics
$BL(\varepsilon)$	Beta-Liouville distribution with parameter ε
$\theta_{jk}/\varepsilon \sim BL(\varepsilon)$	θ_{jk}/ε drawn from $BL(\varepsilon)$
$\varphi_{kv}/\zeta \sim BL(\zeta)$	φ_{kv}/ζ drawn from $BL(\zeta)$
$Mult(\theta_{jk})$	Multinomial distribution with parameter (θ_{jk})
$z_{ik}/\theta_{jk} \sim Mult(\theta_{jk})$	z_{ik}/θ_{jk} drawn from Multinomial(θ_{jk})
$x_i/z_{ik}, \varphi_{kv} \sim Mult(\varphi_{z_{ik}})$	$x_i/z_{ik}, \varphi_{kv}$ drawn from Multinomial$(\varphi_{z_{ik}})$
ψ_{ijk}	Responsability of the component k for the word x_{ij} in document j
$\mathscr{F}(\psi_{ijk}, \theta, \varphi)$	EM lower bound to the log likelihood
$\mathscr{L}(\psi_{ijk}, \theta, \varphi)$	EM lower bound for MAP (maximum a posteriori)
$\mathscr{N}_{-ij}$	Expected Count excluding z_{ij}

5.3 Proposed Approach

In this chapter, we propose a standard EM algorithm for MAP estimation of LBLA topic model. We show that it is an alternative to the CVB algorithm [12]. Moreover, we show that the complexity of the VB approach (when characterizing dependency between latent variables and parameters) ultimately led to the implementation of our simple and standard EM algorithm for MAP estimation: modeling dependence in latent topic models is a way of characterizing accurate parameter estimation from the work in [12]. However, the variational method in the collapsed space can be extremely complex despite its flexibility. We demonstrate that in spite of the simplicity of the proposed EM algorithm for MAP-LBLA, it is implicitly connected to the CVB inference. Furthermore, we cover the generative equation of the MAP-LBLA that allows us to formulate through a coordinate ascent framework the EM-based batch and online algorithms for MAP-LBLA. The accuracy in the expectations also depends on the proposed unnormalized representation.

5.3.1 Modeling dependency between hidden variables

One of the central themes in CVB inference is the possibility to reach accurate parameters estimates by relaxing the independence assumption in mean-field variational methods. This relaxation introduces dependency between latent variables and models parameters. To be more specific, from Table 5.1, with the LBLA hyperparameters ε, ζ, and hidden variables Z, θ, and φ, let's consider the case where we lower bound the log marginal likelihood $\log p(X|\varepsilon,\zeta)$ using variational distributions q:

$$\begin{aligned}\log p(X|\varepsilon,\zeta) &= \log \int_\theta \int_\varphi \sum_Z p(X,Z,\theta,\varphi|\varepsilon,\zeta) d\theta d\varphi \\ &= \log \mathbb{E}_{q(\theta,\varphi,Z)} \left(\frac{p(X,Z,\theta,\varphi|\varepsilon,\zeta)}{q(\theta,\varphi,Z)}\right) \\ &\geq \mathbb{E}_{q(\theta,\varphi,Z)} \log \left(\frac{p(X,Z,\theta,\varphi|\varepsilon,\zeta)}{q(\theta,\varphi,Z)}\right)\end{aligned}$$

This is also equivalent to:

$$\log p(X|\varepsilon,\zeta) \geq \mathbb{E}_{q(\theta,\varphi,Z)} \log(p(X,Z,\theta,\varphi|\varepsilon,\zeta)) - \mathbb{E}_{q(\theta,\varphi,Z)} \log(q(\theta,\varphi,Z))$$

such that:

$$\log p(X|\varepsilon,\zeta) = \mathscr{F}(q,\theta,\varphi,Z) + KL(q(\theta,\varphi,Z)||p(\theta,\varphi,Z|X,\varepsilon,\zeta)) \tag{163}$$

where:

$$\log p(X|\varepsilon,\zeta) \geq \mathscr{F}(q,\theta,\varphi,Z) \tag{164}$$

In the joint space, the variational distribution $q(\theta,\varphi,Z)$ using the mean-field variational is:

$$q(\theta,\varphi,Z) = q(\theta)q(\varphi)q(Z) \tag{165}$$

The variational distribution in (165) characterizes the independence assumption in standard VB inference. In the collapsed space, the variational distribution in (166) follows dependency between latent variables and parameters:

$$q(\theta,\varphi,Z) = q(\theta,\varphi|Z)q(Z) \tag{166}$$

Using the lower bound, we reach the maximum at $q(\theta,\varphi|Z) = p(\theta,\varphi|X,Z)$ where the functional $\mathscr{F}$ (lower bound) now becomes $\mathscr{F}(q,Z)$. From the work in [89, 12], we obtain:

$$\mathscr{F}(q,Z) = \mathbb{E}_{q(Z)} \log(p(X,Z|\varepsilon,\zeta)) - \mathbb{E}_{q(Z)} \log(q(Z))$$

and $\log q(Z_j) = \mathbb{E}_{i\neq j} q(Z) \log(p(X,Z|\varepsilon,\zeta)) + C$ with C being a constant. It leads to:

$$q(Z_j) = \frac{\exp \mathbb{E}_{i\neq j} q(Z) \log(p(X,Z|\varepsilon,\zeta))}{\sum_z \exp \mathbb{E}_{i\neq j} q(Z) \log(p(X,Z|\varepsilon,\zeta))} \tag{167}$$

which is also equivalent to:

$$q(Z_j = k) = \frac{\exp\{\mathbb{E}_{q(Z_{-j})}[\log p(X, Z_{-j}, Z_j = k|\varepsilon,\zeta)]\}}{\sum_{i=1}^K \exp \mathbb{E}_{q(Z_{-j})} \log(p(X, Z_{-j}, Z_j = i|\varepsilon,\zeta))} \tag{168}$$

where $q(Z_j)$ is the update equation for CVB algorithm as illustrated in [12], [80], and [107]. This update equation is really complex and requires the Gaussian approximation along with second order Taylor expansion. The CVB when modeling dependence structure makes the joint variational distributions for the parameters conditioned on the latent variables in (166).

5.3.1.1 The space of parameters and MAP-LBLA

We marginalize out the latent variables, leaving the model (LBLA) parameters. This is a reverse setting of the collapsed representation that integrates out the parameters. A way of modeling dependency (between hidden variables) in MAP estimation is to make the multinomial variational distribution conditioned on the parameters as shown in (169). In this section, we show that using variational methods makes the MAP update equation extremely complex as well; which ultimately leads to a much simpler method using standard EM algorithm. We have the following variational joint distribution:

$$q(\theta, \varphi, Z) = q(\theta, \varphi) q(Z|\theta, \varphi) \tag{169}$$

Here, we get the maximum when $q(Z|\theta, \varphi) = p(Z|\theta, \varphi, X)$ leading to a lower bound functional:

$$\mathscr{F}(q, \theta, \varphi) \quad = \quad \mathbb{E}_{q(\theta,\varphi)} \log p(X, \varphi, \theta|\varepsilon, \zeta) \quad - \quad \mathbb{E}_{q(\theta,\varphi)} \log q(\theta, \varphi)$$

$$\begin{aligned}\log q(Z|\varphi, \theta) &= \mathbb{E}_{q(\varphi,\theta)} \log p(X, Z, \theta, \varphi|\varepsilon, \zeta) + C \\ &\propto \mathbb{E}_{q(\varphi,\theta)} \left[\log(p(X, Z|\theta, \varphi) p(\theta, \varphi|\varepsilon, \zeta))\right] \\ &\propto \mathbb{E}_{q(\varphi,\theta)} \left[\log(p(X, Z|\theta, \varphi) p(\theta|\varepsilon) p(\varphi|\zeta))\right]\end{aligned}$$

We obtain the following update equations for MAP-LBLA:

$$\begin{aligned}\log q(Z|\varphi, \theta) \propto \mathbb{E}_{q(\varphi,\theta)} &\left[\log p(X|Z, \varphi) + \log p(Z|\theta)\right] \\ &+ \mathbb{E}_{q(\varphi,\theta)} \left[\log p(\theta|\varepsilon) + \log p(\varphi|\zeta)\right]\end{aligned} \tag{170}$$

$$\log q(\varphi, \theta) = \mathbb{E}_{q(Z)} \left[\log p(X, Z, \theta, \varphi|\varepsilon, \zeta)\right] + C$$

$$\begin{aligned}\log q(\varphi, \theta) \propto \mathbb{E}_{q(Z)} &\left[\log p(X|Z, \varphi) + \log p(Z|\theta)\right] \\ &+ \mathbb{E}_{q(Z)} \left[\log p(\theta|\varepsilon) + \log p(\varphi|\zeta)\right]\end{aligned} \tag{171}$$

With the Jensen's inequality, providing a lower bound to the log marginal likelihood function $p(X|\varepsilon, \zeta)$ in [12] makes the variational update equation in (171) for MAP intractable because of the coupling between the corpus and document parameters. Even the posterior variational distribution for latent Z in (170) is intractable due to the same coupling. Using the same Jensen's inequality approach, we therefore propose a lower bound to the log likelihood function instead. Then, we derive the MAP lower bound from the log likelihood's lower bound by adding the log of the priors distributions to the log likelihood's lower bound.

5.3.2 Unnormalized parameterization

The stochastic variational inference randomly draws a data point (a word or a document) and then learns its local parameters to update the global parameters following a natural gradient update approach [1, 65]. Let's suppose, for instance, that we are sampling one document at a time. Following the stochastic variational method at document-level, we compute the noisy estimate of the natural gradient of the objective function corrresponding to $\mathscr{D}$ copies of document j which are then used to update the global parameters. As we observe, to allow $\mathscr{D}$ copies of the objective function (ELBO), we take the corpus-wide terms [65], [167] in the variational lower bound of a single document j and normalize them by $\mathscr{D}$ (the total number of documents in the corpus) so that lower bound becomes:

$$\mathscr{L} = \sum_j \mathscr{L}_j = \mathbb{E}_j[\mathscr{D}\mathscr{L}_j] \tag{172}$$

where $\mathscr{D}\mathscr{L}_j$ is the variational lower bound (ELBO) with $\mathscr{D}$ copies of document j.
Similarly, in MAP estimate as we follow this time a stochastic framework at word-level, we need to operate on unnormalized parameterization in order to compute unnormalized expected sufficient statistics during the stochastic expectation step as in MAP-LDA. This is because in online EM algorithm as proposed in [168], the likelihood function and the sufficient statistics are normalized by the total number of words $\mathscr{W}$ in the corpus. Using $\mathscr{W}$ copies of the proposed EM lower bound for each word leads to an unnormalized expected sufficient statistics during E-step and provides the appropriate scale between the normalized ML estimates and the prior distribution that summarizes the posterior probability of the parameters. This shows the correspondence between the proposed approach for MAP where we estimate sufficient statistics within unnormalized representation and the stochastic variational inferences as they use noisy estimates of natural gradient of the ELBO to update the global parameters.

We compute the unnormalized expected sufficient statistics as MAP estimates for the parameters using online averages as alternatives. While performing in unnormalized parameterization of LDA, one of the advantages is the fact that MAP-LDA's update equation and the one for CVB0 (zero order approximation of LDA) are analytically identical if we adjust their hyperparameters by one [4]. This ultimately connects the CVB0-LDA to MAP-LDA and stochastic CVB0-LDA (SCVB0-LDA) to online EM for MAP-LDA, and it introduces the EM statistics and responsibilities to CVB0 statistics and variational distributions (responsibilities). This connection allows the SCVB0-LDA as unnormalized stochastic MAP-LDA with minibatch of size one scheme when assessing one data point at a time (from its recursive update equation). In this chapter, we are also performing in unnormalized parameterization of LBLA where we hope to show its connection to the collapsed space representation. The connection originates from the fact that both MAP and SCVB0 operate on unnormalized parameterization of the LDA. Furthermore, the SCVB0-LDA's update equation is also similar to that of MAP-LDA [13]. We implement a MAP-LBLA estimation with stochasticity at word-level that is connected to SCVB0-LDA inference. In MAP, from the hidden variables, we marginalize out the latent variables from the corpus while leaving only the parameters. On the other hand, in CVB inference, we integrate out the parameters from the hidden variables.

5.3.3 Generative process of the MAP-LBLA

LDA [3, 5] is generally a three level hierarchical model. The corpus level includes the global topics and their hyperparameters and document hyperparameters. The document level is characterized by the topic proportions and finally, the word level includes the topic assignments and the words [5]. By marginalizing out the parameters, we get a two level hierarchical LDA (corpus to document and document to word). As based on the LDA architecture, LBLA in this condition also follows a two level topic model, and as a result, generates documents within the MAP framework as:

```
Choose a global topic $\varphi_k|\zeta \sim \mathrm{BL}(\zeta)$ where $k \in \{1, ..., K\}$
  For each document $j$
    Choose the topic proportion $\theta_j|\varepsilon \sim \mathrm{BL}\ (\varepsilon)$
     For $i \in \{1, 2, ..., \mathscr{W}\}$ in document $j$
      Choose word $x_i|\theta_j, \varphi_{1:K} \sim \mathrm{Mult}\Big(\sum_{i=1}^{K} \theta_{ji}\varphi_i\Big)$
```

In this chapter, the variables x_i, w_i, and v_i could be used interchangeably to denote a word in the vocabulary.

It is noteworthy that this two level hierarchical topic model is different from mixture of unigrams and PLSA or PLSI because its multinomial parameters are drawn from prior distributions (the existence of priors to smooth the multinomials). Without the priors, our LBLA and LDA topic models could have been reduced to PLSI or mixtures of unigrams [4]. Table 5.1 summarizes the relevant variables for the MAP-LBLA topic model.

5.3.4 The Two-Level LBLA Topic Mixture Model

In general from the hidden variables and observed data, the three-level generative equation is:

$$p(X, Z, \theta, \varphi|\varepsilon, \zeta) = \prod_{k=1}^{K} p(\varphi_k|\zeta) \prod_{j=1}^{\mathscr{D}} p(\theta_j|\varepsilon) \prod_{i=1}^{N} p(z_{ij}|\theta_j)p(x_{ij}|\varphi, z_{ij}) \tag{173}$$

We then compute the joint posterior distribution:

$$p(Z, \theta, \varphi|X, \varepsilon, \zeta) = \frac{p(X, Z, \theta, \varphi|\varepsilon, \zeta)}{p(X|\varepsilon, \zeta)}$$

When we marginalize out the latent variables, the two-level LBLA posterior distribution becomes:

$$p(\theta, \varphi|X, \varepsilon, \zeta) = \frac{p(X, \theta, \varphi|\varepsilon, \zeta)}{p(X|\varepsilon, \zeta)} \propto p(X, \theta, \varphi|\varepsilon, \zeta) \tag{174}$$

where

$$p(X, \theta, \varphi|\varepsilon, \zeta) = \sum_{Z} p(X, Z, \theta, \varphi|\varepsilon, \zeta) \tag{175}$$

$$= \sum_{Z} p(X, Z|\theta, \varphi)p(\theta, \varphi|\epsilon, \zeta) \tag{176}$$

with

$$\sum p(X, Z|\theta, \varphi)p(\theta, \varphi|\epsilon, \zeta) = p(\theta|\varepsilon)p(\varphi|\zeta) \sum_{Z} p(X, Z|\theta, \varphi) \tag{177}$$

In our case, $p(\theta_{jk}|\varepsilon)$ and $p(\varphi_k|\zeta)$ are BL priors where $\varepsilon = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ and $\zeta = (\lambda_{k1}, ..., \lambda_{kV}, \lambda, \eta)$ are their respective parameters. For instance, the document BL priors $p(\theta_j|\varepsilon)$ is defined as:

$$p(\theta_j|\varepsilon) = BL(\alpha_1, ..., \alpha_K, \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)\Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}$$

$$\times \prod_{k=1}^{K} \frac{\theta_{jk}^{\alpha_k - 1}}{\Gamma\left(\alpha_k\right)} \left(\sum_{k=1}^{K} \theta_{jk}\right)^{\alpha - \sum_{k=1}^{K} \alpha_k} \left(1 - \sum_{k=1}^{K} \theta_{jk}\right)^{\beta - 1} \tag{178}$$

To show the sufficient statistics and natural parameters of the BL priors for the corpus and documents parameters, we represent them in exponential family form using for instance $p(\theta_d|\epsilon) = \exp\{\log p(\theta_j|\varepsilon)\}$ as we also show below.

$$\begin{aligned} p(\theta_d|\varepsilon) = \exp\{ & \left(\sum_{k=1}^{K}(\alpha_k - 1)\log\theta_{jk}\right) \\ & + \left(\alpha - \sum_{k=1}^{K}\alpha_k\right)\log\left(\sum_{k=1}^{K}\theta_{jk}\right) \\ & + (\beta - 1)\log\left(1 - \sum_{k=1}^{K}\theta_{jk}\right) + \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) \\ & + \log(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) - \sum_{k=1}^{K}\log\Gamma(\alpha_k)\} \end{aligned} \quad (179)$$

From (178) and (179), we use similar steps for the corpus BL prior $p(\varphi_k|\zeta)$.

We define the joint distribution $p(X,\theta,\varphi|\varepsilon,\zeta)$ such that: $p(X,\theta,\varphi|\varepsilon,\zeta) = p(X|\theta,\varphi)p(\theta|\varepsilon)p(\varphi|\zeta)$

$$p(X,\theta,\varphi|\varepsilon,\zeta) = \sum_Z p(X,Z,\theta,\varphi|\varepsilon,\zeta) \quad (180)$$

$$= \left(\sum_Z p(X,Z|\theta,\varphi)\right)p(\theta|\varepsilon)p(\varphi|\zeta) \quad (181)$$

$$= \left(\sum_Z p(X|Z,\varphi)p(Z|\theta)\right)p(\theta|\varepsilon)p(\varphi|\zeta) \quad (182)$$

From (180) to (182), we saw that when the latent variables are marginalized out, the three-level LBLA topic model is reduced to a two-level LBLA model similar to LDA.

The distribution $p(X|\varphi,\theta) = \sum_Z p(X|Z,\varphi)p(Z|\theta)$ is reminiscent of a mixture topic model (PLSI or mixture of unigrams) [5]. Given priors information, we can define:

$$p(X|\varepsilon,\zeta) = \int\int p(\theta|\varepsilon)p(\varphi|\zeta)\left(\prod_{i=1}^{N}p(x_i|\theta,\varphi)\right)d\theta d\varphi \quad (183)$$

This marginal distribution (183) of a document is a (continuous) mixture distribution whose mixture weights are $(p(\theta|\varepsilon) \times p(\varphi|\zeta))$ and components $p(x_i|\theta,\varphi)$ [5]. This illustrates the flexibility of the LBLA due to the prior information compared to PLSI and mixture of unigrams.

5.3.5 The EM lower bound for the MAP-LBLA

We first define the log likelihood $\log P(X|\theta,\varphi)$ as:

$$\log P(X|\theta,\varphi) = \log\sum_Z p(X,Z|\theta,\varphi) \quad (184)$$

We introduce the distribution $q(Z)$ over the latent variables Z. Instead of log marginal distribution, we provide an EM lower bound to the log likelihood which allows us to include the prior information in the EM lower bound for MAP-LBLA in (193).

$$\begin{aligned} \log p(X|\theta,\varphi) &= \mathscr{F}(q,\theta,\varphi) + KL(q||p) && (185)\\ &\geq \mathscr{F}(q,\theta,\varphi) && (186) \end{aligned}$$

with $KL(q||p) \geq 0$

$$\mathscr{F}(q,\theta,\varphi) = \sum_{Z} q(Z) \log \frac{p(X,Z|\theta,\varphi)}{q(Z)} \tag{187}$$

$$KL(q||p) = -\sum_{Z} q(Z) \log \frac{p(Z|X,\theta,\varphi)}{q(Z)} \tag{188}$$

From the definition of the $KL(q||p)$, if $q(Z) = p(Z|X,\theta^0,\varphi^0)$
then $KL(q||p) = 0$, then we have:

$$\mathscr{F}(q,\theta,\varphi) = \sum_{Z} q(Z) \log p(X,Z|\theta,\varphi) - \sum_{Z} q(Z) \log q(Z) \tag{189}$$

$$\begin{aligned} \mathscr{F}(q,\theta,\varphi) = \sum_{Z} p(Z|X,\theta^0,\varphi^0) \log p(X,Z|\theta,\varphi)& \\ - \sum_{Z} p(Z|X,\theta^0,\varphi^0) \log p(Z|X,\theta^0,\varphi^0)& \quad (190) \end{aligned}$$

$$\begin{aligned} \mathscr{F}(q,\theta,\varphi) &= Q(\theta,\varphi,\theta^0,\varphi^0) + C && (191)\\ &= \sum_{Z} p(Z|X,\theta^0,\varphi^0) \log p(X,Z|\theta,\varphi) && (192) \end{aligned}$$

The functional $\mathscr{F}$ represents the standard EM lower bound for ML estimation (MLE) as illustrated in Table 5.1. Now using Bayes' theorem, we can derive an EM lower bound for MAP-LBLA:

$$\begin{aligned} \log p(\theta,\varphi|X) &= \log p(\theta,\varphi,X) - \log p(X)\\ &= \log p(X|\theta,\varphi) + \log p(\theta,\varphi) - \log p(X)\\ &= \mathscr{F}(q,\theta,\varphi) + KL(q||p) + \log p(\theta,\varphi) + C\\ &\geq \mathscr{F}(q,\theta,\varphi) + \log p(\theta,\varphi) + C\\ &\geq \mathscr{F}(q,\theta,\varphi) + \log p(\theta) + \log p(\varphi) + C \end{aligned}$$

Here $KL(q||p) \geq 0$ and $\log p(X)$ is a constant C. Since $q = q(Z) = p(Z|X,\theta^0,\varphi^0) = \psi_{ijk}$ which is our EM responsibility vector, similar to a variational responsibility, then the EM lower bound for MAP is:

$$\mathscr{L}(\psi_{ijk},\theta,\varphi|\varepsilon,\zeta) = \mathscr{F}(\psi_{ijk},\theta,\varphi) + \log p(\theta|\varepsilon) + \log p(\varphi|\zeta) \tag{193}$$

This shows that at the E-step, the MAP lower bound will be identical or reduced to the MLE one if we compute the latent ψ_{ijk} and then the M-step will require both the MLE lower bound and the priors information to estimate the parameters [89]. We have in our case a topic mixture model where its parameters are drawn from their respective conjugate priors.

We showed that when $q(Z) = p(Z|X, \theta^0, \varphi^0) = \psi_{ijk}$ which is the complete conditional distribution of the latent variables given the samples and model parameters, the variational case and the standard mixture model technique coincide. Below, from (194) to (199), we show the MAP steps for its point estimate from its EM lower bound with LBLA.

$$\begin{aligned}\mathscr{L}(\psi_{ijk}, \theta, \varphi) \propto & \left(\sum_k \psi_{ijk} \sum_{i,j,v} \log p(X_i|Z_{ij}, \varphi_{kv}) p(Z_{ij}|\theta_{jk}) \right) \\ & + \left(\sum_{j,k} \log p(\theta_{jk}|\varepsilon) + \sum_{k,v} \log p(\varphi_{kv}|\zeta) \right)\end{aligned} \tag{194}$$

From the lower bound in (194) and (265), we derive the coordinate ascent method that is used to compute the model point estimate θ and φ from (224) to (231). Then we formulate the MAP-LBLA update equation as a function of θ and φ using (231), (229), (195), (196), (197), and (198).

$$\psi_{ijk} \propto (\theta_k)(\varphi_k)(\varphi_{k(V+1)}) \tag{195}$$

$$\psi_{ij(K+1)} \propto \left(\theta_{j(K+1)}\right) \tag{196}$$

$$\begin{aligned}\psi_{ijk} \propto & \left[\frac{\left(\mathscr{N}_\theta^{jk} + \alpha_k - 1\right)}{(\sum_k \alpha_k - 1) + \left(\sum_{k=1}^K \mathscr{N}_\theta^{jk}\right)} \right] \\ & \times \left[\frac{\left(\mathscr{N}_\varphi^{v_{ij}k} + \lambda_{kv} - 1\right)}{(\sum_v \lambda_{kv} - 1) + \left(\sum_{v=1}^V \mathscr{N}_\varphi^{v_{ij}k}\right)} \right] \\ & \times \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)(\varphi_{k(V+1)})\end{aligned} \tag{197}$$

such that:

$$\mathbb{U} = \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)(\varphi_{k(V+1)}) \tag{198}$$

with:

$$\begin{cases} 1 - \theta_{d(K+1)} = \sum_{k=1}^K \theta_{dk} < 1 \\ 1 - \varphi_{k(V+1)} = \sum_{v=1}^V \varphi_{kv} < 1 \\ \varphi_{k(V+1)} = 1 - \sum_{v=1}^V \varphi_{kv} < 1 \end{cases} \tag{199}$$

The Beta distributed random variables in (198) make the MAP-LBLA (197) irreducible to MAP-LDA due to the constraints in (199) which prohibit the factor (198) to be equal to one: as a result, the MAP-LBLA and MAP-LDA do not have the same update equation. However, under some conditions, we could observe that MAP-LBLA update equation in (197) is proportional to that of MAP-LDA in [4], [13], and [17] when using the unnormalized representation. In that case, the EM responsibility vector becomes:

$$\psi_{ijk} \propto \left[\frac{\left(\mathscr{N}_\theta^{jk} + \alpha_k - 1\right)}{(\sum_k \alpha_k - 1) + \left(\sum_{k=1}^K \mathscr{N}_\theta^{jk}\right)} \right] \left[\frac{\left(\mathscr{N}_\varphi^{v_{ij}k} + \lambda_{kv} - 1\right)}{(\sum_v \lambda_{kv} - 1) + \left(\sum_{v=1}^V \mathscr{N}_\varphi^{v_{ij}k}\right)} \right] \tag{200}$$

From (200), the EM algorithm for MAP-LBLA could be identified with MAP-LDA. Using the work in [11], we could also notice that BL prior in (201) contains Beta distribution (203) (the generating density function) that is related to the density generator in (202):

$$p(\theta_j|\alpha_1, ..., \alpha_K) = \mathscr{G}(\gamma) \prod_{k=1}^{K} \frac{\theta_{jk}^{\alpha_k - 1}}{\Gamma(\alpha_k)} \tag{201}$$

The density generator $\mathscr{G}(.)$ of BL gives:

$$\mathscr{G}(\gamma) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\gamma^{\sum_{k=1}^{K} \alpha_k - 1}} \mathscr{J}(\gamma) \tag{202}$$

Below is the representation of the Beta distribution in BL given its hyperparameters α and β.

$$\mathscr{J}(\gamma|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \gamma^{\alpha-1}(1-\gamma)^{\beta-1} \tag{203}$$

with $\gamma = \sum_{k=1}^{K} \theta_{jk} < 1$. From (203), we can use (198) and (197) to show that:

$$\left\{ \varphi_{k(V+1)}(1 - \varphi_{k(V+1)}) \to \textstyle\sum_{v=1}^{V} \varphi_{kv} \sim Beta(2, 2) \right. \tag{204}$$

since $\mathscr{J}(\gamma|2, 2) \propto \gamma(1 - \gamma)$ for $\gamma = \sum_{v=1}^{V} \varphi_{kv}$
Then, we identify the Beta parameters from (204) and (203):

$$\lambda = 2 \qquad\qquad \eta = 2 \tag{205}$$

for the corpus BL prior. From (195) to (197), we can observe that $\mathscr{J}(\gamma|2, 2) \propto \gamma(1 - \gamma)$ for $\gamma = \sum_{v=1}^{V} \theta_{jk}$ as well; so for the document BL prior we have:

$$\alpha = 2 \qquad\qquad \beta = 2 \tag{206}$$

As we identify the hyperparameters of the generating density function or the Beta distribution (203), the corpus BL is then defined as BL$(\lambda_{k1}, ..., \lambda_{kV}, 2, 2)$ while the document BL is still BL$(\alpha_1, ..., \alpha_K, 2, 2)$ from (201), (202), (203), (205), and (206). Importantly, during initializations, the only unknown hyperparameters in the MAP-LBLA are the Liouville distribution document parameters $(\alpha_k)_{k=1}^{K}$ and Liouville corpus parameters $(\lambda_{kv})_{v=1}^{V}$. The EM lower bound to MAP estimation therefore simplifies the LBLA structure which has been reduced from a three-level hierarchical model to two levels. This ultimately suggests that while the formulation of MAP-LBLA in (197) is proportional to the update equation in (200) which bears some ressemblance with MAP-LDA [4], [17], [13], we can primarily identify (200) as a Liouville family distribution that turns out to be proportional to the Dirichlet. Since the Liouville family distribution of the second kind is proportional to Dirichlet, then both their update equations under a topic modeling framework would be proportional. This is the case because in (200) by proportionality, the Beta prior defined in (204) acts as a uniform prior. As previously mentioned, when considering proportionality, the MAP-LBLA's update equations could be equivalent to those from MAP-LDA. Instead of using EM statistics in a form of $\mathscr{N}_\varphi$, $\mathscr{N}_\theta^j$, and $\mathscr{N}_Z$, we could also represent the EM algorithm for LBLA point estimates in terms of unnormalized counts of the EM responsibilities as shown below:

$$\theta_{jk} \propto \left(\mathscr{N}_\theta^{jk} + \alpha_k - 1 \right) (1 - \theta_{j(K+1)}) \tag{207}$$

$$\propto \left(\mathscr{N}_\theta^{jk} + \alpha_k - 1 \right) = \sum_i \psi_{ijk} + \alpha_k - 1 \tag{208}$$

$$\varphi_{kv} \propto \left(\mathscr{N}_{\varphi}^{v_{ij}k} + \lambda_{kv} - 1\right)(1 - \varphi_{(V+1)k}) \tag{209}$$

$$\propto \left(\mathscr{N}_{\varphi}^{v_{ij}k} + \lambda_{kv} - 1\right) = \sum_{ij} \psi_{ijk} + \lambda_{kv} - 1 \tag{210}$$

where the EM statistics for LBLA are:

$$\begin{cases} \mathscr{N}_{\theta}^{jk} = \sum_i \psi_{ijk} \\ \mathscr{N}_{\varphi}^{kv} = \sum_j \psi_{ijk} \\ \mathscr{N}_{Z}^{k} = \sum_{ij} \psi_{ijk} \end{cases} \tag{211}$$

We just showed that with unnormalized count, the LBLA using EM for MAP, and LDA share similar parameters and EM statistics in (229), (231), and (211). So we combine unnormalized count method to parameterization to connect the MAP estimation to other inferences such as stochastic variational inference for the LDA architecture. We will show that our proposed approach could be in alignment with the work in [4].

The batch algorithm for MAP using EM for LBLA (BEM-LBLA) follows (229), (231), and (197). It requires an extensive amount of memory because it stores on each word, in the corpus, an EM responsibility vector. It constantly needs access to all the available data at every iteration before providing an update which is not efficient. We first aim for a fast batch method (similar to CVB0) from which we can build a stochastic EM algorithm for MAP estimation.

5.3.6 Fast Batch Algorithm for EM-LBLA

It is mainly a refined version of the original batch EM algorithm for MAP-LBLA. For time and memory complexity, it is directly faster and provides a good performance over the original batch EM because it excludes the current posterior from its sufficient statistics. It excludes current value of the responsabilty for the word x. The counts in (212) are then used for batch processing. In the collapsed space, it is equivalent to CVB0. We summarize its expected counts or EM sufficient statistics as:

$$\begin{cases} \mathscr{N}_{\theta-ij}^{jk} = \sum_{-i} \psi_{ijk} \\ \mathscr{N}_{\varphi-ij}^{v_{ij}k} = \sum_{-j} \psi_{ijk} \\ \mathscr{N}_{Z-ij}^{k} = \sum_{-(i,j)} \psi_{ijk} \end{cases} \tag{212}$$

where the responsability update equation is defined as:

$$\begin{aligned} \psi_{ijk} \propto & \left[\frac{\left(\mathscr{N}_{\theta-v_{ij}}^{jk} + \alpha_k - 1\right)}{\left(\sum_{k=1}^{K} \alpha_k - 1\right) + \left(\sum_{k=1}^{K} \mathscr{N}_{\theta-v_{ij}}^{jk}\right)} \right] \\ & \times \left[\frac{\left(\mathscr{N}_{\varphi-v_{ij}}^{v_{ij}k} + \lambda_{kv} - 1\right)}{\left(\sum_v \lambda_{kv} - 1\right) + \left(\sum_{v=1}^{V} \mathscr{N}_{\varphi-v_{ij}}^{v_{ij}k}\right)} \right] \\ & \times \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)\left(\varphi_{k(V+1)}\right) \end{aligned} \tag{213}$$

with ij meaning the ith word in document j; ijk is the ith word in document j in topic k; i also represents the number of latent variables in the document j. From the work in [4] and [13], it shows that SCVB0-LDA is equivalent to MAP-LDA through unnormalized parameterization because the MAP-LDA update equation is identical to the SCVB0 except that their hyperparameters must be offset by one. This equivalent relationship between the MAP-LDA and SCVB0-LDA characterizes the similarity between the EM statistics and responsibilities with CVB0-LDA statistics and variational responsibilities (distributions). The SCVB0-LDA is the unnormalized MAP-LDA using standard EM algorithm. Because it implements a stochastic method at word-level using the variational distribution as a local parameter, the SCVB0-LDA is the stochastic unnormalized MAP-LDA using minibatch scheme of size one. In the EM algorithm, the MAP estimates unnormalized expected sufficient statistics to scale properly its prior distributions for a normalized likelihood function. In this chapter and using online stochastic CVB0 for LBLA that we previously proposed in [107], we can observe that there is no equivalent relationship between the currently proposed MAP-LBLA and the SCVB0-LBLA [107] as their respective update equations are different. This is in contrast to the MAP-LDA and the SCVB0-LDA that share some equivalent relationship as shown in [13], [17]. However, our EM-based MAP-LBLA shares some equivalent relationship with SCVB0-LDA under unnormalized parameterization in (197) from [13]. As the current MAP-LBLA's update equation is proportional to the one in MAP-LDA, consequently, it is connected to SCVB0-LDA. With unnormalized representation, the EM-based MAP-LBLA associates its EM statistics and responsibilities to CVB0-LDA statistics and responsibilities: the EM algorithm for MAP-LBLA therefore operates on unnormalized parameterization of LDA.

We have just illustrated that both MAP-LBLA and MAP-LDA operate on unnormalized parameterization of LDA. Therefore, the SCVB0-LDA could characterize a MAP estimation for LBLA as well. Connecting the MAP-LBLA to CVB0-LDA and SCVB0-LDA will help in providing an alternative to a model selection as we show in sub-section 5.3.8. The proposed MAP-LBLA ultimately optimizes an EM lower bound on the posterior probability of the parameters using EM responsibilities and EM statistics.

5.3.7 Stochastic EM algorithm for MAP-LBLA model

We call this method SEM-LBLA which stands for stochastic EM algorithm for MAP estimation for LBLA topic model. This method ultimately does not require all the available samples for an update as in the original batch. It follows a stochastic technique within a minibatch scheme that also refines the fast batch in sub-section 5.3.6. The standard batch method is slow. This approach provides two update equations for its global parameters: we have the update equation for the document global parameter and the one for the global topics.

In our proposed method, SEM-LBLA operates as follow: In minibatch, SEM-LBLA accesses one word x in a corpus (a random and uniform draw), then from that sample, it updates its parameters. In the E-step, it computes unnormalized expected sufficient statistics and evaluates the EM responsibility ψ_{ij} associated to the word $x = x_{ij}$. In the M-step, it evaluates the intermediate global parameters (topics in terms of expected counts) in the corpus as it optimizes the EM lower bound. To do that, it creates $\mathscr{W}$ copies of the intermediate global parameters associated to x in the minibatch and then average them using $\mathscr{V}$. The average estimate of the intermediate global parameters in a minibatch (of size $\mathscr{V}$) scheme using $\mathscr{W}$ copies is generally given as:

$$\hat{\mathscr{N}_{\varphi}} = \frac{\mathscr{W}}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \mathscr{A}^{(i,j)} \tag{214}$$

where $\mathscr{A}^{(i,j)}$ is the word-topic expected count matrix. It is a $K \times V$ matrix such that the vth column contains the responsibility

$$\psi_{ij} = \sum_k \psi_{ijk} \tag{215}$$

So for $\mathscr{V} = 1$, we have a minibatch of size one. When $\mathscr{V} > 1$ we have a standard minibatch scheme which draws a subset of samples from the corpus at each iteration. When $\mathscr{W} = \mathscr{V}$ and $\kappa = 0$, we have a batch EM for MAP. However, for a minibatch of size one, the estimate accesses $\mathscr{W}$ copies of the distribution on word x; so the estimate becomes:

$$\hat{\mathscr{N}_\varphi} = \mathscr{W} \sum_{v_{ij} \in \mathscr{V}} \mathscr{A}^{(i,j)} = \mathscr{W} \psi_{ij}[v_{ij} = v] \tag{216}$$

This simply counts the number of times the word v appears in the corpus (of size $\mathscr{W}$) as the global intermediate estimate for selecting a random word v in the corpus. Then, this intermediate global parameter estimate is then used to update the global topic parameter as shown in:

$$\mathscr{N}_\varphi[t+1] = (1 - \rho_t)\mathscr{N}_\varphi[t] + \rho_t \hat{\mathscr{N}_\varphi} \tag{217}$$

where $\rho_t = (\tau_0 + t)^{-\kappa}$ is the step size. The variable τ_0 is the number of minibatches (predefined), t is the minibatch index, and $\kappa \in (0.5 \;\; 1]$ is provided by the users. Similarly, in the document j, the random and uniform draw of a word in a corpus creates an intermediate global estimate of $\hat{\mathscr{N}_\theta^j} = \mathscr{W}_j \psi_{ij}$ (for $\mathscr{V} = 1$) leading to an update equation of:

$$\mathscr{N}_\theta^j[t+1] = (1 - \rho_t)\mathscr{N}_\theta^j[t] + \rho_t \hat{\mathscr{N}_\theta^j} \tag{218}$$

When $\mathscr{V} > 1$, we use:

$$\begin{cases} \hat{\mathscr{N}_\theta^j} = \frac{\mathscr{W}_j}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \psi_{ij} \\ \hat{\mathscr{N}_Z} = \frac{\mathscr{W}}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \psi_{ij} \end{cases} \tag{219}$$

We also estimate the expected count $\hat{\mathscr{N}_Z} = \mathscr{W} \psi_{ij}$ (when $\mathscr{V} = 1$) and then summarize its online average equation in the following:

$$\mathscr{N}_Z[t+1] = (1 - \rho_t)\mathscr{N}_Z[t] + \rho_t \hat{\mathscr{N}_Z} \tag{220}$$

From [13], the SEM-LBLA will converge to the stationary point of the MAP objective function. This is because:
$0 < \rho_t \leq 1 \; \forall \; t$ and $\sum_t^\infty \rho_t = \infty$ and $\lim_{\infty} \rho_t = 0$. These expectations explain the EM statistics for the MAP-LBLA along with the responsibility vector ψ_{ijk}. The parameter estimates at M-step are identical to the expected sufficient statistics from E-step.

5.3.8 Model selection: Small samples, number of topics, and vocabulary size under MAP-LBLA

The MAP favors small samples as it can regularize better ML estimates with its prior information. Because it can perform well on small datasets, we expect it to offer a much improved performance when using for instance a minibatch processing (stochastic method) compared to full batch methods. For extremely large samples, the MAP estimate will be close to the posterior means (unbiased estimator [162], [157]). However, it will require expensive computational resources [157]. Large samples can cause an increase in the number of parameters, especially the number of topics and vocabulary size. Increasing the number of topics in a finite, parametric topic mixture model is not ideal because such setting automatically increases the search space for the optimal number of topics and vocabulary size [157], [80]. To efficiently reduce the searching space for model selection, we propose a performance of our MAP algorithm using small samples size along with small number of topics and possibly small vocabulary size as well [17]. From previous sections, we showed that our model, the MAP-LBLA, under unnormalized parameterization is equivalent to SCVB0 which uses CVB0 (the zero order approximation of CVB) as a fast batch method [13]. The CVB0 itself could be restrictive for large scale processing because of its memory requirement problems at every iteration [13], [157]. It led to a stochastic CVB0 or SCVB0. The work in [157] showed that CVB0 favors a small set of topics because when the hypothesis grows, the CVB0 is unable to find a global optimum as it often gets stuck in local maxima [157], [1]. The SCVBO uses its stochasticity to escape local optima [157], [1]. SCVB0 can operate in large scale applications (Big Data), but it has no ability in parameter streaming especially when the vocabulary size and number of topics increase in topic-word matrix. To solve this problem the SCVB0 could operate on a small number of topics, and small minibatches, possibly using minibatch of size one.

Since the MAP-LBLA is connected to SCVB0-LDA through MAP-LDA, it can therefore carry such implementation to allow both large scale data and parameter streaming. This explains our decision to operate on small number of topics and samples sizes for MAP-LBLA topic model. In terms of EM algorithm, under unnormalized counts, the MAP-LDA and MAP-LBLA have similar update equations. We can use these characteristics to assess a model selection for our LBLA model through LDA since MAP-LDA and SCVB0 have identical update equations with only their hyperparameters offset by one [4, 13]. In other words, from SCVB0-LDA to MAP-LBLA, the MAP update equation only adds negative one on its hyperparameters. The SCVB0-LDA is therefore the unnormalized representation of online EM for both MAP-LDA and MAP-LBLA. However, from analysis, SCVB0 uses CVB0 as a fast batch method in a stochastic optimization. Despite its use of large memory, the CVB0 could outperform the unbiased estimator CGS when the number of topics is low [157]. As a deterministic method, this allows it to converge faster than any other inferences. Finally, for instance, in multi-label framework [157], when only one sample is required, the CVB0 outperforms both the CGS and its unbiased estimator. Since SCVB0 is a stochastic version of CVB0, it carries all the advantages of CVB0 while improving the memory requirement of CVB0 for large scale data processing.

In topic modeling, time and memory complexities are functions of the number of topics and the size of the vocabulary [17, 12]. When the size of global parameters increases, especially the word-topic expected count matrix of size $K \times V$, a model selection that efficiently reduces the variables K and V ultimately improves time and memory complexities. Such model selection scheme would implicitly provide an efficient setting for a parameter streaming. We would like to consider improving solutions provided to SCVB0-LDA in model selection and data management problems with our proposed EM-based MAP-LBLA topic model. This is because our approach is connected to SCVB0 through the MAP-LDA: from the literature, as SCVB0 (with a minibatch scheme that processes one sample at a time [13]) is equivalent to a stochastic unnormalized MAP for LDA, we can therefore set a minibatch of size one for MAP-LBLA as we suggested it earlier in sub-section 5.3.7 to accommodate data and parameter streaming. This constitutes a direct alternative to the work in [17] that uses a dynamic scheduling approach based on residuals including a buffer mechanism that provides an alternative to model selection which also improves both time and memory complexities. The approach in [17] first reduces the number of topics and vocabulary size in a parametric finite topic mixture model using LDA. Their buffering technique makes it easy to transfer data between computer's memory and external storages that carry the load (massive data including word-topics expected count matrices). This finally leads to a parameter streaming that fixes the problem of big topic modeling in large scale applications.

Our proposed alternative to model selection using minibatch scheme of size one or reasonable minibatch sizes is in agreement with the core method that is implemented in [17]. Though, ours is more simpler and also allows us to process documents with almost infinite vocabularies: a minibatch of size one ultimately fixes the problem of vocabulary. This is equivalent to processing or updating only the vth column of the $K \times V$ word-topic matrix while the corpus expected count increases by one anytime we access a new vocabulary, for instance. We can summarize our contribution as follows: ultimately, with our scheme supporting a small number of topics and a minibatch method of size one, there is no need for a buffer of size $K \times V$ to connect to external storages. This facilitates flow of data and parameter. In fact, the buffering scheme would have required us to probably implement two buffers: one for the global topic matrix and one for the document parameter (documents expected count matrix), and use both simultaneously in inferences within a stochastic framework at word level which defines the topic and document parameters as global parameters. In contrast to the method in [17] which follows a stochasticity at document-level, our approach does not discard the document parameter after one look. It updates both the corpus (topics) and document parameters. We only used the connection between the MAP-LBLA and MAP-LDA and their equivalent relationship within the collapsed variational Bayes inferences to suggest for an improved alternative to model selection for MAP in order to handle both large scale data and parameter streaming. Our method is not computationally expensive when we compare it to the work in [17] that supports expensive dynamic scheduling and a buffering methods.

In our proposed method, despite being stochastic, we also prioritize accurate estimates over extremely fast methods that could miss important processing steps and negatively affect overall results. For a regular minibatch (255), with reasonable small samples, we can use the proposed $|K| \leq 150$ as almost similar to the setting in [157], and for a minibatch scheme of size one (216), we can set $|K| = 10$ for every word as in [17]. We combine both processes in our framework where we use regular minibatch when the parameters and data are manageable or we switch to a minibatch of size one for extremely large vocabulary size in the data and parameters.

5.4 Experimental results and settings

5.4.1 Datasets

We consider three challenging text document datasets. These collections are: ENRON dataset, NIPS text documents, and KOS data as shown in Table 6.6. ENRON dataset has total corpus of $\mathscr{D} = 39861$ documents. With a vocabulary size $V = 28102$, it provides a total of $\mathscr{W} = 6400000$ words. The NIPS text documents represent a collection from scientific papers from the proceedings of NIPS database. It has a corpus around 2484 papers. The corpus contains $\mathscr{D} = 1740$ documents for a total vocabulary size $V = 12419$. It also carries a total of $\mathscr{W} = 2166029$ words and $M = 836644$ unique word-document pairs. The KOS collection is from the report blog website (online). It has a total of $\mathscr{D} = 3430$ documents, a vocabulary size of $V = 6909$, and a total of $\mathscr{W} = 467714$ words and $M = 360664$ unique word-document pairs.

Table 5.2: Text document datasets

	$\mathscr{D}$train	$\mathscr{D}$test	$\mathscr{W}$	V	$\mathscr{D}$
NIPS	1256	419	2166029	12419	1675
KOS	2573	857	467714	6909	3430
ENRON	29896	9965	6400000	28102	39861

5.4.2 Implementation

This is a stochastic EM algorithm for MAP estimation using the LBLA topic model. As we perform a stochastic at word-level method in our proposed approach, we have two global parameters to estimate instead of one global parameter as in case of a stochasticity at document level. Our global parameters include the topic-word parameters and the document parameters. We estimate these parameters in terms of unnormalized expected counts which define our EM statistics for the stochastic EM-based LBLA model for MAP estimation. The M-step optimizes the EM lower bound with respect to the parameters while the E-step provides the unnormalized expected sufficient statistics as we use here exponential family distributions. The proposed approach requires initializations on the hyperparameters. We usually set them randomly. However, for the BL hyperparameters, we also provide initializations as follows: For BL prior on the document multinomial parameter, we choose $\alpha_{jk} = \frac{1}{k}$ where $k \in \{1, 2, ..., K\}$ to characterize asymmetric BL prior. We set $\alpha_j = 2$ based on (206), and we choose α_{jk} such that $\alpha_j - \sum_{k=1}^{K} \alpha_{jk} \neq 0$. Then, we choose $\beta_j = 2$. For the BL on the corpus multinomial parameter, we are setting values for λ_{kv} with $v \in \{1, 2, ..., V\}$ (similar to the document BL) and $\lambda = 2$ (where $\lambda - \sum_{v=1}^{V} \lambda_{kv} \neq 0$) and $\eta = 2$ from (205) for every k. We use a stochasticity at word-level where we randomly sample one word at a time from which we estimate its EM responsibility vector (local parameter) ψ_{ijk} that allows us to obtain estimates of the model parameters in terms of expected counts.

We implement a minibatch method of size one to process one sample at a time. We use regular minibatch (multiple samples) when the parameters and data are manageable or we can also switch to a minibatch of size one for extremely large vocabulary size in the data. This illustrates the flexibility of our framework to large scale applications. At convergence, the global parameters are approximated as point estimates. The method still favors much smaller batch size so that the prior regularizes estimates. We set the minibatch sizes as: $\mathscr{V} = \{10, 40, 60, 80, 100\}$. The set of topics is: $K = \{10, 20, 40, 60, 80, 100, 120, 150\}$. We provided a learning rate ρ_t at iteration t such that:

$$\rho_t = (t + \tau_0)^{-\kappa} \tag{221}$$

The forgetting rate $\kappa \in (0.5, 1]$ actively controls how quickly previously estimated data are forgotten, during successive iterations. With EM algorithm, we can always reach a local optimum of the EM lower bound of the posterior distribution of the parameters. We maintain $\tau_0 = 1$ and $\kappa = 0.7$.

5.4.2.1 Evaluation method using perplexity

Each of the three datasets selected for this experiment went through similar process. In each dataset (a collection of text documents), we randomly divide the data into training and testing sets. We compute the corpus parameters φ during the training phase. Then, in the test document, we randomly divide it into a ratio of 90% and 10% as each subset contains word tokens. As we fix φ, we estimate the document topic proportions θ on the 90% of the test set and then calculate the predictive perplexity on the rest 10% of the subset using (222) in [153].

A low value of the predictive perplexity or a high predictive log likelihood suggest a better model.

$$perplexity = \exp\left\{-\frac{\sum_{i,j} x_{ij} \log[\sum_k \psi_{ijk}]}{\sum_{ij} x_{ij}}\right\} \tag{222}$$

The variables x_{ij} and $[\sum_k \psi_{ijk}]$ represent the data and responsibility at 10%, respectively. We compute the responsibility vector ψ_{ijk} using (197) from our EM statistics while φ is maintained fixed.

5.4.2.2 Time and memory complexities

The proposed online EM based-MAP-LBLA has similar time and memory complexity to LDA topic model in general [12, 153]. Especially, the work in [153] has provided an extensive detail on LDA's time and memory complexities. Though, the main difference between the LDA and LBLA's time and memory complexities is the flexibility of the BL that allows the model to perform many tasks: its covariance structure offers possibility to model selection easier than the one in LDA when analyzing topic structure based on probability masses (topic proportions) associated to global topics in LBLA. The LDA has no ability to topic correlation analysis as we mentioned earlier. Therefore, our model is much faster because it can handle more tasks than LDA including performing topic correlation analysis; all these tasks within the same time of LDA. This suggests that online EM based-MAP-LBLA is faster at performing each task and therefore has a much improved time complexity compared to its LDA counterpart per task. Furthermore, with flexible priors such as BL, it means we do not need too much samples including the number of topics to achieve better estimates as the MAP improves and regularizes our point estimates.

Since in topic modeling, time and memory complexities are functions of the parameters such as the number of topics K the size of the vocabulary V, and the size of the dataset $\mathscr{D}$ [12], [153], when K, V, and $\mathscr{D}$ become extremely small, they can significantly improve the memory requirement (with stochastic method) and the time complexity. The possibility in our case to carry extremely small samples makes it a better approach over the LDA in terms of time and memory complexities. It also makes online method efficient over batch techniques.

5.4.3 Results

The use of prior distributions for MAP estimation makes PLSA and mixture of unigrams unfit for comparison because the work in [3, 5] even used the simple symmetric LDA to show the limitations of the PLSA and mixture of unigrams as they lack prior information. In this experiment, we mainly focus on topic models that could characterize a Bayesian framework. We compare the performance of our LBLA topic model directly to LDA for

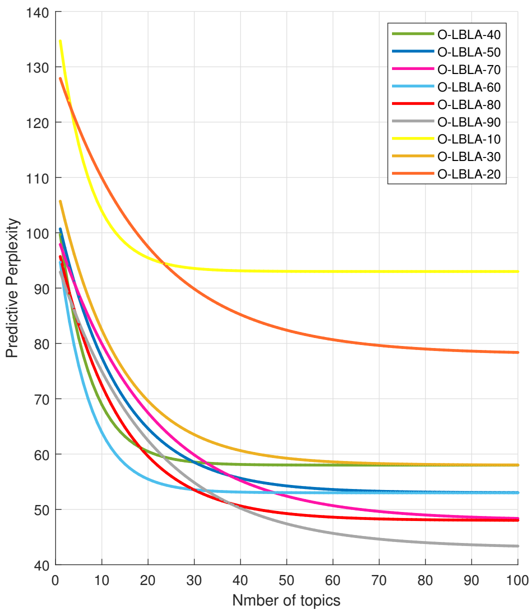
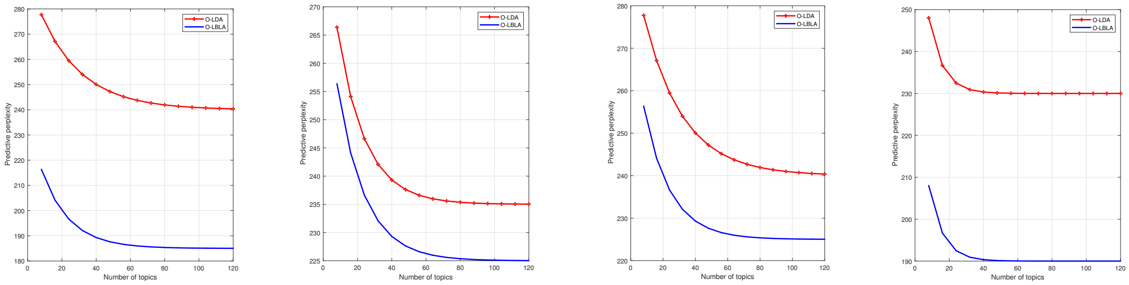


Figure 5.1: Online MAP-LBLA and NIPS batch sizes



(a) $\mathscr{V} = 20$

(b) $\mathscr{V} = 40$

(c) $\mathscr{V} = 60$

(d) $\mathscr{V} = 80$

Figure 5.2: Online EM-based MAP-LBLA vs. online EM-based MAP-LDA at different minibatch sizes (NIPS dataset)

MAP estimation. We then use the predictive perplexity to evaluate the online EM algorithm for MAP-LBLA and MAP-LDA under a variety of situations: in each dataset, we monitor the influence of the number of topics and batch size in the predictive perplexity. In each dataset we observe that online EM for MAP-LBLA is faster than online EM for MAP-LDA because of its ability to summarize relevant topics faster than symmetric LDA. Importantly, we observe that the predictive perplexity favors a small number of topics as we assess the first topic values to which the perplexity remains constant while being at its lowest values. The online EM for MAP-LBLA constantly outperforms online EM for MAP-LDA in terms of predictive perplexity. Figs. 5.2, 5.4, and 5.6 show the performance of the LBLA over the symmetric LDA in each of our proposed datasets. The flexibility of the BL prior in LBLA also plays a central role in the predictive distributions and perplexity: a topic model in general has a fixed multinomial distribution as likelihood function. Its robustness relies on

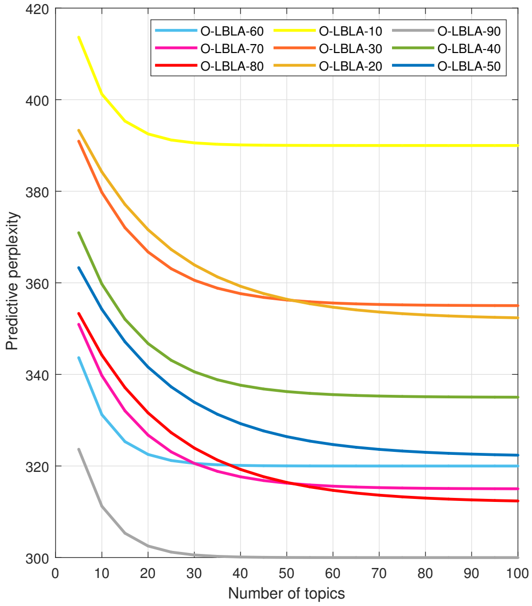
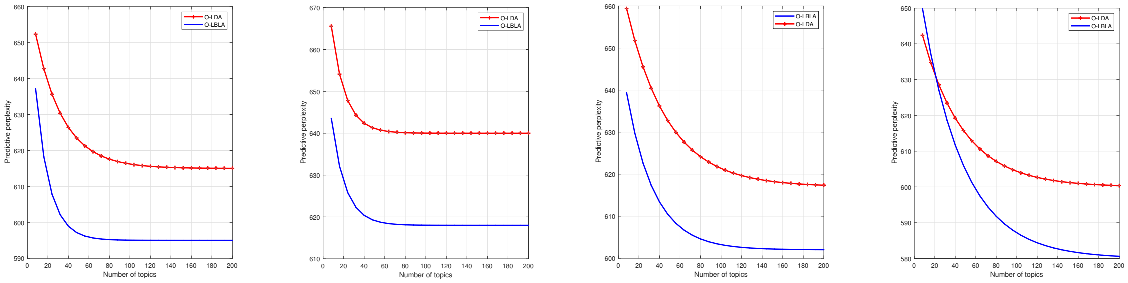


Figure 5.3: Online MAP-LBLA and KOS batch sizes



(a) $\mathscr{V} = 20$ (b) $\mathscr{V} = 40$ (c) $\mathscr{V} = 60$ (d) $\mathscr{V} = 80$

Figure 5.4: Online EM-based MAP-LBLA vs. online EM-based MAP-LDA at different minibatch sizes (KOS dataset)

the choice of priors such as BL. The symmetric prior with a uniform base measure does not offer a variability in the set of topics while the asymmetric BL prior provides heterogeneity in the topics that speeds up the search for most relevant topics. In addition, the use of uniform priors such as symmetric Dir, while it simplifies computation, reduces the MAP framework to MLE. Within the MAP-LBLA topic models, we also observe that providing a reasonable batch size ultimately enhances the predictive performance in our datasets from Figs. 5.1, 5.3, and 5.5. This is because a reasonable size of samples could benefit from the contribution of prior information In this case the MAP could act as regularizer through the prior for small sample sizes. These characteristics in the proposed approach improve point estimates and contributes to a much robust perplexity framework. It is also important to mention that in many occasions, the predictive perplexity of the MAP-LDA is almost close to that of the MAP-LBLA as shown in Figs 5.2b, 5.6a, and 5.6c. This could be

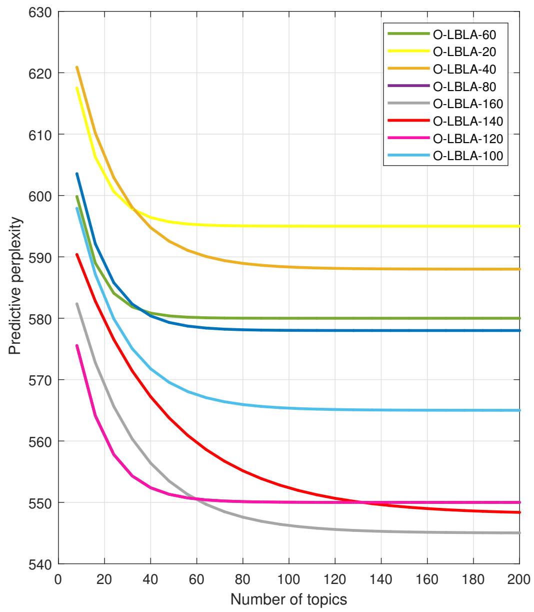
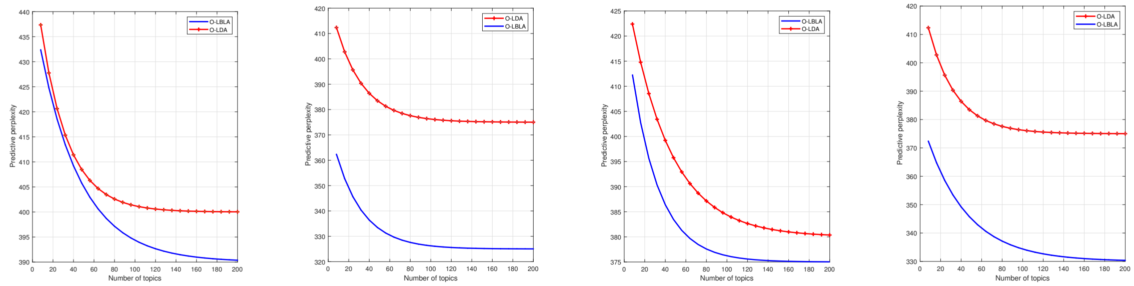


Figure 5.5: Online MAP-LBLA and ENRON batch sizes



(a) $\mathscr{V} = 20$

(b) $\mathscr{V} = 40$

(c) $\mathscr{V} = 60$

(d) $\mathscr{V} = 80$

Figure 5.6: Online EM-based MAP-LBLA vs. online EM-based MAP-LDA at different minibatch sizes (ENRON dataset)

explained by the hyperparameter setting in LBLA. The LBLA is a generalization of LDA which means under some conditions (hyperparameter initialization) the LBLA could be reduced to LDA topic model. The MAP-LBLA favoring a small number of topics and a relatively reasonable batch size show its equivalent relationship with CVB0 that also favors small number of topics [157].

5.5 Conclusion

In this chapter, for parameter estimation in topic modeling, we provide an alternative to the collapsed variational Bayes and collapsed Gibbs inferences by proposing a simple MAP estimation technique based on standard EM algorithm. The method optimizes an EM lower bound on the posterior distribution of the parameters in the M-step. In the E-step, it updates exponential family sufficient statistics using online averages. Our main parameters are the unnormalized expected counts (EM statistics) that summarize the MAP-LBLA's update equation. The CVB and CGS, the collapsed space inferences, marginalize out the parameters while leaving the latent variables. On the other hand, the MAP estimation method integrates out the latent variables leaving only the parameters. It also reduces the three-level hierarchical structure in topic models to two levels in the hierarchy. We implement the MAP-LBLA using online EM algorithm and then compare its performance (predictive perplexity) against the MAP-LDA that is with equipped symmetric Dir. We show that the update equation of MAP-LBLA could be proportional to that of MAP-LDA.

The MAP-LDA is connected to CVB0 because they have identical update equations with only their hyperparameters adjusted or offset by one. The CVB0 favors a small number of topics. The stochastic CVB0 (SCVB0) allows large scale data modeling but could not handle parameter streaming due to the size of vocabulary and number of topics as they increase in large scale processing. The MAP-LBLA (which is connected to MAP-LDA) aims to improve the capability of SCVB0-LDA that has an equivalent relationship with MAP-LDA: under unnormalized parameterization, the SCVB0-LDA is equivalent to MAP-LDA. Furthermore, using reasonable samples sizes in the minibatch scheme ultimately fixes the problem related to large parameter matrices especially the word-topic expected count matrix during inferences. We manage the data and parameter streaming by creating a framework where we use regular minibatch when the parameters and data are manageable or we switch to a minibatch of size one for extremely large vocabulary sizes in the data. Because the number of topics and vocabulary size are reduced in this way, the memory and time complexities are much improved in the proposed approach. We also think that the efficiency in the predictive perplexities is due to the flexibility of the BL prior in LBLA compared to the Dir distribution in LDA. Its ability to model dependency between documents through topic correlation characterizes a much robust compression algorithm and predictive models. It is still important to recognize that in general, one of the problems in parametric finite topic mixture models is the parameters initializations, especially the number of topics. In addition, these models seem to have a much reduced hypothesis space that do not allow them to cope with extremely large number of topics.

For future work, we could investigate the performance of the topic model when using other flexible conjugate priors such as generalized Dirichlet based on hyperparameter estimation. Similarly, we could also implement non conjugate priors using for instance logistic normal distributions. Another alternative to finite mixture topic models would be to implement nonparametric models where datasets ultimately choose their underlined components (number of topics) by themselves.

Appendix

We formulate the EM lower bound for MAP-LBLA where the priors $\mathscr{L}(\psi_{ijk}, \theta, \varphi)$ are BL distributions.

$$
\begin{aligned}
\mathscr{L}(\psi_{ijk}, \theta, \varphi) \propto \sum_{k,i,j,v} & \psi_{ijk} \log \varphi_{kv} + \psi_{ijk} \log \theta_{jk} \\
& + \psi_{ij(K+1)} \log(\theta_{j(K+1)}) + \psi_{ijk} \log(\varphi_{k(V+1)}) \\
& + \{ \left(\sum_{j,k} (\alpha_k - 1) \log \theta_{jk} \right) + \left(\alpha - \sum_k \alpha_k \right) \log \left(\sum_{j,k} \theta_{jk} \right) \\
& + (\beta - 1) \log \left(1 - \sum_{j,k} \theta_{jk} \right) + \log \Gamma \left(\sum_{k=1} \alpha_k \right) + \log \Gamma \left(\alpha + \beta \right) \\
& - \log \Gamma(\alpha) - \log \Gamma(\beta) - \sum_k \log \Gamma(\alpha_k) \} \\
& + \{ \left(\sum_{k,v} (\lambda_{kv} - 1) \log \varphi_{kv} \right) + \left(\lambda - \sum_v \lambda_{kv} \right) \log \left(\sum_{k,v} \varphi_{kv} \right) \\
& + (\eta - 1) \log \left(1 - \sum_{k,v} \varphi_{kv} \right) + \log \Gamma \left(\sum_v \lambda_{kv} \right) + \log \Gamma(\lambda + \eta) \\
& - \log \Gamma(\lambda) - \log \Gamma(\eta) - \sum_v \log \Gamma(\lambda_{kv}) \} \qquad (223)
\end{aligned}
$$

We perform a coordinate ascent method to obtain the parameter update equations: we characterize the lower bound associated to each parameter, compute the corresponding derivative and set it equal to zero. We added the Lagrangian term to the lower bound to include the optimizations constraints for the parameters before derivation.

$$
\begin{aligned}
\mathscr{L}(\theta) = \sum_{k,i,j,v} & \psi_{ijk} \left(\log \theta_{jk} \right) + \{ \left(\sum_{j,k} (\alpha_k - 1) \log \theta_{jk} \right) \} \\
& + \left(\alpha - \sum_k \alpha_k \right) \log \left(\sum_{j,k} \theta_{dk} \right) + (\beta - 1) \log \left(1 - \sum_{j,k} \theta_{jk} \right) \\
& + \xi \left(\theta_{j(K+1)} + \sum_{k=1}^{K} \theta_{jk} \right) \qquad (224)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{jk}} \mathscr{L}(\theta) = & \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} + \frac{\alpha - \sum_k \alpha_k}{\sum_{k,j} \theta_{jk}} \\
& + \frac{1 - \beta}{1 - \sum_{k,j} \theta_{jk}} + \xi
\end{aligned} \qquad (225)
$$

Let $\mathscr{T}$ be defined as: $\mathscr{T} = \frac{\alpha - \sum_k \alpha_k}{\sum_{k,d} \theta_{dk}} + \frac{1-\beta}{1 - \sum_{k,j} \theta_{jk}}$, so we can see that $\mathscr{T}$ is not defined when $\sum_{k,j} \theta_{jk} = 0$ and $1 - \sum_{k,j} \theta_{jk} = 0$;

$\sum_{k,j} \theta_{jk} \neq 0$ and $(1 - \sum_{k,j} \theta_{jk}) = \theta_{j(K+1)} \neq 0$, $\mathscr{T} = 0$ means

$$\alpha = \sum_k \alpha_k \qquad\qquad \beta = 1 \tag{226}$$

So we have:

$$\frac{\partial}{\partial \theta_{jk}} \mathscr{L}(\theta) = \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} + \mathscr{T} + \xi \tag{227}$$

Now making the derivative equal to zero gives $\frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} + \mathscr{T} + \xi = 0$ or $\frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} = -\mathscr{T} - \xi$; so $\theta_{dk} = \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{-C - \xi}$ where $\sum_k \theta_{jk} = \sum_k \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{-\mathscr{T} - \xi} = 1 - \theta_{j(K+1)}$; $-\xi = \frac{\sum_k (\sum_i \psi_{ijk} + \alpha_k - 1) + \mathscr{T}(1 - \theta_{j(K+1)})}{1 - \theta_{j(K+1)}}$ $\theta_{jk} = \frac{\sum_i \psi_{ijk} + \alpha_k - 1}{\frac{\sum_k (\sum_i \psi_{ijk} + \alpha_k - 1) + \mathscr{T}(1 - \theta_{j(K+1)})}{1 - \theta_{j(K+1)}} - \mathscr{T}}$ or

$$\theta_{jk} = \frac{\sum_i \psi_{ijk} + \alpha_k - 1}{\frac{\sum_k (\sum_n \psi_{ijk} + \alpha_k - 1) + \mathscr{T}(1 - \theta_{j(K+1)}) - \mathscr{T}(1 - \theta_{j(K+1)})}{1 - \theta_{j(K+1)}}} \tag{228}$$

$\theta_{jk} = \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\frac{\sum_k (\sum_n \psi_{ijk} + \alpha_k - 1)}{1 - \theta_{j(K+1)}}} = \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\sum_k \sum_i \psi_{ijk} + \alpha_k - 1} (1 - \theta_{j(K+1)})$

For $\mathscr{N}_\theta^{jk} = \sum_i \psi_{ijk}$

$$\theta_{jk} = \frac{\left(\mathscr{N}_\theta^{jk} + \alpha_k - 1 \right)}{(\sum_k \alpha_k - 1) + \left(\sum_k \mathscr{N}_\theta^{jk} \right)} (1 - \theta_{j(K+1)}) \tag{229}$$

We have also

$$\begin{aligned} \mathscr{L}(\varphi) = \sum_{k,i,j,v} \psi_{ijk} (\log \varphi_{kv}) &+ \left(\sum_{k,v} (\lambda_{kv} - 1) \log \varphi_{kv} \right) \\ &+ \left(\lambda - \sum_v \lambda_{kv} \right) \log \left(\sum_{k,v} \varphi_{kv} \right) \\ &+ (\eta - 1) \log \left(1 - \sum_{k,v} \varphi_{kv} \right) + \varrho \left(\varphi_{(V+1)k} + \sum_{v=1}^{V} \varphi_{kv} \right) \end{aligned} \tag{230}$$

Similarly for φ_{kv} using $\mathscr{L}(\varphi)$, we have:

$$\varphi_{kv} = \frac{\left(\mathscr{N}_\varphi^{v_{ij}k} + \lambda_{kv} - 1 \right)}{(\sum_v \lambda_{kv} - 1) + \left(\sum_v \mathscr{N}_\varphi^{v_{ij}k} \right)} (1 - \varphi_{k(V+1)}) \tag{231}$$

We define Ω similar to $\mathscr{T}$ as $\Omega = \frac{\lambda - \sum_v \lambda_{kv}}{\sum_{k,v} \varphi_{kv}} + \frac{1 - \eta}{1 - \sum_{k,v} \varphi_{kv}}$ with $\mathscr{N}_\varphi^{vjk} = \sum_{(ij)=v} \psi_{ijk}$ where the ith word is v
with $1 - \theta_{j(K+1)} = \sum_{k=1}^{K} \theta_{jk}$ and $1 - \varphi_{k(V+1)} = \sum_{v=1}^{V} \varphi_{kv}$

Chapter 6

Stochastic Variational Optimization of A Hierarchical Dirichlet Process Latent Beta-Liouville Topic Model

In topic models, collections are organized as documents where they arise as mixtures over latent clusters called topics. A topic is a distribution over the vocabulary. In large scale applications, parametric or finite topic mixture models such as LDA (latent Dirichlet allocation) and its variants are very restrictive in performance due to their reduced hypothesis space. In this chapter, we address the problem related to model selection and sharing ability of topics across multiple documents in standard parametric topic models. We propose as an alternative a BNP (Bayesian nonparametric) topic model where the HDP (hierarchical Dirichlet process) prior models documents topic mixtures through their multinomials on infinite simplex. We therefore propose asymmetric BL (Beta-Liouville) as a diffuse base measure at the corpus level DP (Dirichlet process) over a measurable space. This step illustrates the highly heterogeneous structure in the set of all topics that describes the corpus probability measure. For consistency in posterior inference and predictive distributions, we efficiently characterize random probability measures whose limits are the global and local DPs to approximate the HDP from the stick-breaking's formulation with the GEM (Griffiths-Engen-McCloskey) random variables. Due to the diffuse measure with the BL prior as conjugate to the count data distribution, we obtain an improved version of the standard HDP that is usually based on symmetric Dirichlet (Dir). In addition, to improve coordinate ascent framework while taking advantage of its deterministic nature, our model implements an online optimization method based on stochastic, at document level, variational inference to accommodate fast topic learning when processing large collections of text documents with natural gradient. The high value in the likelihood per document obtained when compared to the performance of its competitors is also consistent with the robustness of our fully asymmetric BL-based HDP. We show that online HDP-LBLA (Latent BL Allocation)'s performance is the asymptote for parametric topic models. The accuracy in the results (improved predictive distributions of the held out) is a product of the model's ability to efficiently characterize dependency between documents (topic correlation) as now they can easily share topics, resulting in a much robust and realistic compression algorithm for information modeling.

6.1 Introduction

In this world of data analytics, the introduction of Bayesian approaches has revolutionized data mining and machine learning techniques for large scale applications. One of the alternatives provided by Bayesian analysis to classical frequentist estimation (e.g. maximum likelihood estimation) remains the use of prior distributions in data modeling to accommodate the likelihood functions [89, 7]. Today, the possibility of applying full Bayesian techniques where we can marginalize over the entire parameters space ultimately opens the door to better prediction rules [42, 9, 18] including the possibility to compare efficiently different models [89]. This concept is being currently applied in topic modeling with the extensive use of probabilities as ways to quantity uncertainty. The main goal is to learn good topics that could be used in applications such as classification and information retrieval (i.e. a robust topic model could be embedded into a search engine). For classification, for instance, the topics learnt in the generative stage could be used in the discriminative stage with powerful classifiers such as SVM (support vector machines).

Topic models in general depend heavily on prior distributions because their likelihood functions are already intrinsically fixed to multinomial distributions, so robust topic models are characterized by the flexibility in the choice of the prior distributions. It is noteworthy that the multinomial distribution carries some limitations that are widely known and documented in machine learning's literature. In text document analysis, a multinomial distribution does not capture very well the phenomenon of word burstiness, as shown in [6, 7]. Furthermore, as it usually operates with count data, a direct frequentist method with multinomial distributions is not efficient because it provides unstable point estimates where unseen events are more likely to get zero probability for highly sparse data. A prior distribution ultimately solves these problems by smoothing out the multinomial. Many applications have therefore favored the Dirichlet as prior distribution to the multinomials [3, 7]. The use of prior information is very necessary to point out the highly dependence of topic models on a Bayesian analysis. Without these priors, the widely known parametric LDA topic model would be reduced to a mixture of unigrams or PLSA (probabilistic latent semantic analysis) which do not offer the advantages of full Bayesian methods [3, 4]. Topic models can be viewed also as compression algorithms that summarize documents into relevant topics. In other words, in topic modeling, documents arise as mixtures of topics where each topic (the latent cluster) is a distribution over the vocabulary. For better compression algorithms, it is natural to characterize an efficient topic modeling's architecture where documents show some level of dependency between them through their topic structure. Documents that share similar topics could be grouped together. This provides a much realistic and natural way of organizing unstructured collections [8]. The current popularity of topic models due to the proliferation of large scale applications (text document modeling and computer vision) has made it necessary to point out that the majority of these topic models are severely challenged by complexities in datasets. To cope with the difficulties, finite dimensional topic models (parametric) have to navigate between conjugate prior distributions [3, 5, 107, 67, 80, 64, 35] and non conjugate priors [8, 21]. Conjugate priors to the multinomials make posteriors in the same form as the priors. They tend to make inferences simple with possible closed-form solutions within the mean-field variational approach. For instance, the Dirichlet is considered conjugate prior to the multinomial distribution. The logistic normal distribution used as non conjugate prior is more sophisticated than conjugate priors; however it makes inference extremely complex and computationally extensive. Logistic normal distribution models the topic proportions to allow full covariance structure on the topic mixture components [8, 169, 21]. Its main contribution was to model correlation between topics as a way of characterizing dependency between documents. Still, to improve results in the standard parametric LDA topic model, symmetric and asymmetric priors have been introduced, and they have shown which combinations enhance topic models performances [9]. For LDA, the standard method requires a use of asymmetric Dirichlet prior for the document parameter and symmetric Dirichlet prior on the corpus parameter leading to AS (asymmetric-symmetric) combination following the work in [9]. We can immediately observe that all these specifications and restrictions ultimately increase the complexity in topic modeling just in the choice of a simple, but flexible prior as we now have to navigate through layers of conjugate and non conjugate priors before assessing flexibility of our model with asymmetric over symmetric structure or maybe efficiently combining these priors for better results [9]. Importantly, the complexity in the choice of priors is highly increased by the ultimate possibility of assessing the appropriate number of latent clusters (topics) that describe the datasets [5, 80, 41, 9]. As in any latent mixture model, model selection [16, 170, 171] is an important subject in parametric and nonparametric topic mixture models. The machine learning's literature has suggested several methods for topic mixture models [16, 170, 171]. The problem with working with finite mixtures is that their reduced hypothesis space prohibits a robust model selection ability. For instance, the deterministic nature of CVBO (zero order approximation of the collapsed variational Bayes) [4] combined with its reduced hypothesis space [172] prevent the model from performing efficiently when there is an increase in the number of topics. The scheme can effectively outperform the unbiased estimator CGS [173] when the number of topics is low. When there is an increase in the number of topics, CVB0 performs poorly. Since working in finite dimensional space with parametric topic model is very challenging due to the choice of prior that can complicate inferences and estimates including problems related to reduced hypothesis space and efficient model selection, we propose a framework where the dataset chooses its own components. In other words, we let the dataset choose its underlined structure instead of making any assumption as often in parametric models. The proposed technique has a much bigger hypothesis space (operating in infinite dimensional space) that allows to cope and accommodate any sort of complexity. Another reason for our proposed alternative is the widely use of LDA and its Dirichlet prior in many applications. Since with Dirichlet prior, the topic components are independent [8], it does not offer a more natural and realistic way of exploring unstructured (large) collections of observed data where correlation might highly occur.

Compared to standard finite mixture models, topic models generalize the concept of finite mixtures as each observation arises from multiple draws from a mixture model. As a result, they are also called mixed membership models [1]. Parametric finite dimensional topic models extensively use symmetric Dirichlet priors in LDA. The problem with the symmetric Dirichlet is that it forces the topics to exhibit same frequency (topics are equally common) [173, 1]. As a result, this could make model selection very difficult for parametric topic models. In nonparametric topic models, the literature has presented several methods with Bayesian nonparametric priors such as the DP (Dirichlet process) [174, 175] and HDP (hierarchical Dirichlet process) [14], the two parameter poisson-Dirichlet process also called Pitman-Yor process (PYP), [176] and its hierarchical extensions. The DP and HDP have prediction rules that introduce the Chinese restaurant process (CRP) and the Chinese restaurant franchise (CRF), respectively [14]. As the PYP and HPYP generalize the DP and HDP, respectively, they carry prediction rules that generalize the CRP and CRF, respectively [15]. While the HPYP is used in natural language processing [177], [178] and image segmentation [179], the HDP prior is widely used in topic modeling where the symmetric Dirichlet serves as a base probability measure at the top level DP which is often referred to as the HDP-LDA topic model. We are aware that providing the best priors will result in a better topic model. Through an efficient prior, and due to the limitation of parametric finite dimensional topic models, we aim for a nonparametric setting that could serve as an alternative to model selection while characterizing the sharing ability of clusters (topics) between documents. The DP is the nonparametric prior when extending standard parametric mixture model whereas the HDP is the appropriate nonparametric prior for topic models. We therefore propose as alternative to the standard HDP-LDA, a stochastic Bayesian nonparametric (BNP) technique that implements a variant of HDP prior where the base distribution at the top level DP (Dirichlet process) is the asymmetric Beta-Liouville distribution. In contrast to the standard methods that implement symmetric Dirichlet prior at the top level DP, we formulate the BL prior as alternative for more variability and heterogeneity in documents topic structure [173]. The goal is to characterize a detailed topical multi-resolution analysis (where coarser topics interact with finely grained topics) for accurate estimates [180]. By providing the atoms/topics, the BL is the conjugate prior to the data distribution in the documents. This flexibility allows us to carry efficient topic correlation framework within a nonparametric setting. The BL with its versatile covariance structure could be the alternative to the discrete infinite logistic distribution (DILN) that has been proposed in HDP because implementing continuous priors on discrete distributions was hindering performance when topics are not sparse [181]. Since the logistic normal distribution has very complex posterior inference, our proposed topic model with the BL as conjugate prior to the multinomial could be the appropriate nonparametric topic correlation method that not only could offer an alternative to model selection but also allow topic sharing. This will result in a much robust and efficient compression algorithm in topic modeling for information retrieval. With the use of BL, we formulate the proposed approach as asymmetric HDP-LBLA (HDP based on the latent BL allocation) in comparison to the HDP-LDA topic model [15, 1, 167, 182]. One more leverage from our HDP-LBLA is that while sharing the global topics among its documents, the topics could be highly sparse than in HDP-LDA. Therefore, this sparse representation on infinite dimensional space could be useful when selecting most relevant topics. It therefore means, the HDP-LBLA has a very flexible GEM [183] structure at the top and lower level DPs. Finally, we took advantage of the deterministic nature of the proposed approach to implement a fast stochastic variational HDP-LBLA (SV-HDP-LBLA) to accommodate massive collections of documents processing. The consistent high value in the predictive likelihood per word in a document obtained against its competitors shows the robustness of our model in document processing.

Importantly, for an easy implementation of our BNP topic model, we first constructed the parametric SV-LBLA (stochastic variational-LBLA) from which we derive our proposed SV-HDP-LBLA topic model. In our proposed approach, the parametric LBLA is a smoothed topic model where both the document and corpus parameters are drawn from asymmetric Beta-Liouville priors. It is also an alternative to another parametric LBLA variant presented in [64, 35] that is a hybrid between the LBLA and PLSA as it was halfway Bayesian [4]. We characterize a fully Bayesian analysis in our parametric LBLA and nonparametric HDP-LBLA. Three main contributions fundamentally summarize our proposed approach.

- Under standard symmetric LDA, all topics are more likely to exhibit same frequency. We implement a much flexible and sophisticated nonparametric (asymmetric) HDP prior based on BL distribution (the BL generalizes the Dirichlet) which is conjugate to the documents multinomials to enhance variability and heterogeneity of the topics shared among documents. The proposed approach promotes topic correlation.

- With a much improved stochastic variational framework, the heterogeneity in the set of all topics from the global probability measure enhances the GEM structure and contributes to a much faster detection of most relevant topics where the proposed HDP shows that it represents an alternative to model selection. Unlike our nonparametric model, the performance of parametric LDA and its variants including LBLA always rely on the number of topics (which is unknown).
- When marginalizing over the parameters space, the BL-based HDP topic model provides accurate predictive distributions that could enhance perplexities and log likelihood estimates for a better compression algorithm in information retrieval in large scale applications.

This chapter is structured as follows: section 6.2 presents the related work and background while section 6.3 elaborates on our smoothed asymmetric and stochastic LBLA model. Section 6.4 focuses on our proposed stochastic and asymmetric HDP-LBLA model follows by section 6.5 which carries the experiments using text data collections. Finally, section 6.6 provides future works and a conclusion.

6.2 Related work and background

One of the main advantages of topic modeling is in fact the simplicity of its architecture because the likelihood function has been kept fixed to a multinomial distribution as we operate with count data. In generative parametric topic models, as often characterized in posterior inferences, prior distributions play a central role in smoothing the multinomials [6, 4, 58]. Different priors lead to different results in point estimate. This observation has led to the introduction of a wide variety of priors within finite parametric topic models because the goal is to generalize the LDA topic model so that it operates on a variety of data (images, videos and text data). This generalization introduces some complexities and challenges when characterizing useful properties and notions such as topic correlation and dependency between documents through latent clusters (topics). Choosing the right prior for the related multinomials parameters that could respect such properties while enhancing predictive distributions and perplexities has become one of the central themes in efficient topic modeling. The reason behind is to help also in model selection besides the concept of sharing topics between documents which can be described as modeling some level of dependency between corpus documents. In that regard, topic modeling literature has witnessed a competition between conjugate priors (to the multinomials) and non conjugate priors. Conjugate priors [3, 5] are often considered appropriate in inference. The Dirichlet [3] is a widely used conjugate prior in topic models. Non conjugate priors have been essentially introduced for a direct topic correlation framework while characterizing possible connections between documents through their topics. The logistic normal distribution and its variants [8, 181, 21] remain important non conjugate priors for the multinomial. However, the choice of such priors often leads to very complex inferences. Recently, another property has been introduced to the priors adding more fuel to the old battle between conjugate and non conjugate priors: the new battle is now between symmetric (the base measure is fixed to a uniform distribution) and asymmetric priors (base measure fixed to a non uniform distribution) [9].

In LDA, authors in [9] suggested that the combination of an asymmetric prior on the document parameter and a symmetric prior on the corpus parameter, resulting in asymmetric-symmetric (AS) LDA provides better results than any other combination: this has become the standard setting when it comes to choosing priors for multinomials [9] in finite parametric models. In BNP topic models with HDP, maintaining the AS structure is implicitly equivalent to using symmetric Dirichlet priors as base distribution at the corpus level DP whose global atoms are then shared among documents at the lower level DP. For instance, in [16, 167], the truncated versions of the HDP-LDA with a symmetric Dirichlet on the corpus parameter have been observed to implicitly generate asymmetric-symmetric (AS) structure. So, many authors encouraged the widely use of symmetric Dirichlet priors within nonparametric settings as in [182, 14, 1]. In addition, Mallet, for instance, automatically follows an asymmetric-symmetric (AS) LDA structure [184] because it is a truncated version of HDP-LDA with a finite symmetric Dirichlet that truncates a GEM (Griffiths, Engen and McCloskey) [180].

Recently, fully asymmetric priors characterizing the asymmetric-asymmetric (AA) LDA started to emerge in BNP using HDP-LDA to increase flexibility of topics components. This is because researches have recently concluded that topic models hyperparameters can affect the number of topics, so it is unsafe to immediately fix these hyperparameters. The NP-LDA (nonparametric-LDA) is a fully nonparametric asymmetric variant of the HDP-LDA with a truncated GEM prior [180]. This topic model is based on the PYP process, and it is an example of the superiority and flexibility of AA-structure as it outperforms the standard truncated HDP-LDA and variants. It outperforms these models in perplexity, therefore, in predictive distributions [180]. Though, currently, very few authors successfully implemented asymmetric-asymmetric (AA) priors [67, 80, 107, 163] in parametric topic modeling framework. In addition, only a small number of authors have worked on AA structure within nonparametric setting. For instance, the work in [180, 173] supported asymmetric Dirichlet prior at the top level DP and was able to exhibit successfully AA LDA structure instead of the standard AS LDA in [9]. Asymmetric and symmetric Dirichlets have been constantly part of Bayesian inferences. For instance, while authors in [173] applied a CGS (collapsed Gibbs sampler) using asymmetric Dirichlet prior, the work in [180] showed that the truncated versions of the HDP-LDA could not outperform the fully asymmetric nonparametric topic model (NP-LDA) as it is equipped with PYP (Pitman-Yor process) prior on the document and corpus parameters. The hyperparameters have GEM priors and form mean vectors drawn from asymmetric priors. Nonparametric LDA-based topic models that follow AS framework are less robust compared to the NP-LDA with its AA structure. The work in [180] found that AA version of HDP easily outperforms AS methods.

Within hyperparameters analysis, prior distributions in empirical Bayes estimation have successfully demonstrated in the literature their influences in topic models inferences. The work in [4, 180, 9] illustrates the impact of hyperparameters [4] in inferences and model selection [180, 9]. For instance, in model selection, authors in [180] showed that hyperparameters have direct effect on the number of topics. Through the work in [4] which connected hyperparameter analysis to inferences, the literature has also noticed a variety of BNP topic models with HDP often within variational, MCMC (Markov Chain Monte Carlo) approaches (such as Gibbs sampler and collapsed Gibbs sampler), and hybrid methods that bridge the gap between MCMC and variational inferences. An example is the collapsed variational Bayes (CVB) and its variants including online methods with CVB-HDP in[16].

In inferences, the work in [185] implemented a truncation-free method for variational approaches using hierarchical BNP model with DP mixtures as a direct extension to the PYP (Pitman-Yor Process) [176], NCRP (Nested Chinese Restaurant process) [186], [187], and IBP (Indian Buffet Process), [188, 189]. However, these methods including the work in [185] were all outperformed by the work in [173] that implemented an improved version of the direct assignment sampler of [14] and [190]. The scheme in [173] is an online sparse collapsed hybrid variational method that performs better than the Gibbs sampler when using large number of topics. It outperforms the SMF-HDP (stochastic mean-field variational HDP) model [167], and it is an extension to [191] that is based on fast Gibbs sampling using high level of sparsity in hybrid parametric LDA including [68]. The method improves the Chinese restaurant process (CRP) representation. Despite all these flexibilities, the vast majority of these inferences promote the use of Dirichlet priors. Under LDA, with the Dirichlet prior, the topic components are independent, a setting that negatively affects a natural way of visualizing, analyzing, and organizing unstructured collections of data where correlation and information (topics) sharing become essential [8]. Sharing the topics between documents is one the main advantages of nonparametric models using HDP as it provides a better way of grouping documents that are similar leading to a much robust and efficient compression algorithm with HDPs. While Gibbs sampler is an unbiased estimator that has been combined in many hybrid models [12, 173, 70, 172, 16, 170, 171], variational techniques are deterministic and flexible for fast convergence. MCMC approaches are often slow and could not take advantage of fast online schemes (streaming) for massive data processing as variational.

The flexibility of priors could be handicapped by the reduced hypothesis space of finite parametric latent topic models. For instance, CVB0 is unable to perform effectively when the number of topics increases; however, it outperforms the unbiased estimator CGS for a much lower number in topics [172]. This suggests that the reduced hypothesis space can negatively affect model selection or makes it very challenging [5, 9, 41, 153, 1, 80] in finite dimensional space. In many occasions this leads to exhaustive methods including cross-validations that may not be efficient [1] in nonparametric setting with large scale applications. Nonparametric topic models have a much bigger hypothesis space as they operate in infinite dimensional space where the dataset can effectively select its underlined components from a set of a countably infinite components.

The work in [170, 192] recently found there is no significant difference between symmetric and asymmetric priors when using Dirichlet on the corpus parameter in LDA: this shows again the inability of the Dirichlet in general in very complex and challenging applications. We decide to implement a variational approach to take advantage of its deterministic nature while performing an efficient stochastic learning on massive collections of data. We also decide to use a flexible conjugate prior to the data distribution that could equally have the same performance of the logistic normal distribution in infinite dimensional space, a framework proposed in [181] to accommodate topic correlation modeling. Instead of the standard LDA and its symmetric Dirichlet priors that often dominate topic modeling, we propose an alternative with asymmetric BL based-HDP prior for the multinomials in our BNP method. Importantly, the choice of asymmetric BL is an alternative to non conjugate priors such as logistic normal distribution [8] and discrete infinite logistic normal distribution [181]. This suggests that our variational posteriors are simple and in closed-form, unlike the ones in DILN of [181]. In the proposed approach, the BL acts as a diffuse base measure at the top level DP and as a result offers a variant of HDP prior that is different from standard HDP-LDA. With the BL, a global probability measure (a draw from the top level DP with probability one) is highly heterogeneous and holds the set of all topics that are shared among documents at the lower level DPs. The Sethuraman's stick-breaking representations of the probability measures from our two level HDP improve predictive distributions as we marginalize out the parameters. Finally, the stochastic implementation learns rapidly topics components from large corpora (collections). We called it the stochastic variational Bayes using HDP for LBLA topic model (SV-HDP-LBLA).

6.3 Stochastic variational approach using smoothed and fully asymmetric LBLA

6.3.1 Motivations

Our proposed HDP-LBLA topic model is the extended nonparametric version of the fully asymmetric parametric LBLA topic model. The parametric LBLA naturally is a generalization of the parametric LDA due to the BL generalizing the Dirichlet (Dir) prior in LDA. Before actual implementation of our proposed HDP-LBLA in section 6.4, we are first implementing a version of LBLA topic model which will facilitate the transition to HDP-LBLA. In this section, besides the literature review, we are providing some direct motivational steps that have contributed to the implementation of this particular parametric LBLA and proposed HDP-LBLA. We found this step necessary in order to show the core differences between parametric LBLA and its nonparametric counterpart within our variational stochastic inference with predictive scheme for both models.

The original parametric LBLA proposed in [64, 35] is a hybrid topic model between a variational Bayes (VB) and a PLSA's (Probabilistic latent semantic analysis) maximum likelihood method [4]. The work in [64, 35] implements variational posteriors on document parameters; however, the corpus parameter is maintained fixed and estimated through MLE instead as no prior is placed on it. This makes variational coordinate ascent framework inefficient because of the lack of a fully Bayesian analysis in the method which also penalizes its ability to make efficient prediction on unseen documents. Placing a meaningful prior (such as BL in our proposed approach) on the corpus parameter smooths out the corpus multinomials [4, 6, 7] during the training phase. Symmetric priors are widely used in parametric topic modeling many times just for convenience and simplicity in the models [4]. With symmetric priors, the topics are equally common [1] in the set. For instance, with symmetric Dir priors in LDA, we expect all the topics to have more likely the same frequency [173]. Symmetric priors could make model selection very difficult when all topics become equiprobable as they prevent from assessing relevant topics. Despite the flexibility of asymmetric parametric LBLA over parametric LDA, all these parametric models do not have the big hypothesis space of nonparametric topic models including hierarchical topic models to handle model selection and the sharing ability of topics between documents.

Compared to our previous work with LBLA topic model [107] where we implemented a deterministic method with CVB inference, we currently provide another alternative by using a variational method instead of our previous methods with CVB inference. This is due to the efficiency of VB over the complexity of the CVB inference in large scale applications. The CVB is a robust deterministic algorithm; however, it can be slow to reach converge due to its MCMC scheme that utilizes a collapsed Gibbs sampler to compute full dense probability distributions over each token in order to approximate the variational posterior distribution of the topic assignments [157]. The CVB update equation is extremely and computationally expensive, and it requires a second order Taylor expansion as approximation [12, 70, 4] that carries a correction factor of variances. Even the implementation of the zero order information of CVB called CVB0 is also complex and slow due to the fact that it also maintains full dense probability distributions over each word, preventing it from a possible sparsity framework that could speed up the algorithm [157]. These reasons directly motivate our work here where we favor asymmetric priors over symmetric ones. For parametric LBLA topic model, we use asymmetric BL priors as they also generalize the Dirichlet priors; and for our proposed hierarchical nonparametric topic model, we utilize asymmetric HDP prior. Precisely, documents multinomials are drawn from the proposed asymmetric BL-based HDP prior. The HDP-LBLA could be seen as a nonparametric alternative of the parametric topic models such as LBLA and LDA including their parametric variants.

6.3.2 Generative process of LDA and LBLA

LDA [3, 5] is the simplest parametric latent graphical topic model. It assumes that each document exhibits K topics with different proportions. We consider the smoothed LDA topic model where the document and corpus multinomial parameters are drawn from Dirichlet conjugate priors. The generative process is given as:
Draw topics $\varphi_k \sim Dir(\lambda_{k1}, ..., \lambda_{kV})$ for $k \in \{1, 2, ..., K\}$
For each document $d \in \{1, 2, ..., \mathscr{D}\}$:
$a-$Draw topic proportions $\theta_d \sim Dir(\alpha_{d1}, ..., \alpha_{dK})$
$b-$For each word x_{dn} where $n \in \{1, ..., N\}$
i)-Draw topic assignment $z_{dn}|\theta_d \sim Mult(\theta_d)$
ii)-Draw word $x_{dn}|z_{dn}, \varphi_k \sim Mult(\varphi_{z_{dn}})$
where $Mult$ is the multinomial distribution. In LBLA, we replace the Dir for the corpus and documents parameters by $BL(\lambda_{k1}, ..., \lambda_{kV}, \lambda, \eta)$ and $BL(\alpha_1, ..., \alpha_K, \alpha, \beta)$, respectively. Therefore, the hierarchical structure in LDA is also preserved under the LBLA model as well. Nevertheless, in parametric topic models, the LBLA due to its BL prior [107] topic generalizes the LDA.

6.3.3 Asymmetric BL with a general covariance structure

From [11], a vector $\vec{\theta}_d = \{\theta_{d1}, ..., \theta_{dK}\}$ following the BL distribution with parameter $\varepsilon = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ is defined as:

$$p(\vec{\theta}_d|\alpha_1, ..., \alpha_K, \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{k=1}^K \frac{\theta_{dk}^{\alpha_k - 1}}{\Gamma(\alpha_k)} \left(\sum_{k=1}^K \theta_{dk}\right)^{\alpha - \sum_{k=1}^K \alpha_k} \left(1 - \sum_{k=1}^K \theta_{dk}\right)^{\beta - 1} \tag{232}$$

We utilize asymmetric BL priors with non uniform base measures with concentration parameters ε and ζ [9]. Following the work in [7, 11, 64, 35], the expectation of θ_k using BL$(\alpha_1, ..., \alpha_K, \alpha, \beta)$ is:

$$\mathbb{E}[\theta_{dk}|\alpha_k, \alpha, \beta] = \frac{\alpha}{(\alpha+\beta)} \frac{\alpha_k}{\sum_{d=1}^K \alpha_k} \tag{233}$$

$$Var(\theta_{dk}) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \frac{\alpha_k(\alpha_k+1)}{(\sum_{k=1}^K \alpha_k + 1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \frac{\alpha_k^4}{(\sum_{k=1}^K \alpha_k)^4} \tag{234}$$

The BL covariance is defined as:

$$Cov(\theta_{dl}, \theta_{dk}) = \frac{\alpha_l \alpha_k}{\sum_{d=1}^K \alpha_k} \left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)(\sum_{k=1}^K \alpha_k + 1)}\right) - \frac{\alpha^2}{(\alpha+\beta)^2(\sum_{k=1}^K \alpha_k)} \tag{235}$$

$$\mathbb{E}(\log \theta_{dk}|\alpha_k, \alpha, \beta) = \Psi(\alpha) - \Psi(\alpha + \beta) + \Psi(\alpha_k) - \Psi\left(\sum_{d=1}^{K} \alpha_k\right) \tag{236}$$

The vector $\vec{\theta}_d = \{\theta_{d1}, ..., \theta_{dK}\}$ in BL satisfies $\sum_{k=1}^{K} \theta_d < 1$. Therefore, in a K+1-dimensional space $\vec{\theta}$ only carries its first K components.

6.3.4 Variational inference for asymmetric and smoothed LBLA

We compute the lower bound then use it as the objective function to derive a form of a coordinate ascent framework (as we estimate the global variables from the local document context) that implements a stochastic optimization using the natural gradient of the objective function. Given our LBLA model, we compute the following marginal:

$$p(\mathscr{D}|\varepsilon, \zeta) = \int_\theta \int_\varphi \sum_z p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta) d\theta d\varphi \tag{237}$$

$$\begin{aligned} \log p(\mathscr{D}|\varepsilon, \zeta) &= \log \int_\theta \int_\varphi \sum_z p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta) d\theta d\varphi \\ &= \log \int_\theta \int_\varphi \sum_z p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta) \frac{q(\theta, \varphi, z)}{q(\theta, \varphi, z)} d\theta d\varphi \\ &= \log \mathbb{E}_{q(\theta,\varphi,z)} \left[\frac{p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta)}{q(\theta, \varphi, z)}\right] \\ &\geq \mathbb{E}_{q(\theta,\varphi,z)} \left[\log\left(\frac{p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta)}{q(\theta, \varphi, z)}\right)\right] \\ &\geq \mathbb{E}_q \left[\log p(\mathscr{D}, z, \theta, \varphi|\varepsilon, \zeta)\right] - \mathbb{E}_q \left[\log q(\theta, \varphi, z)\right] \\ &= \mathscr{L}(q) \end{aligned} \tag{238}$$

For the stochastic optimization within the variational inference we first need to formulate the objective function (ELBO) in terms of $\mathscr{D}$ copies of document m. This is given in (239).

$$\begin{aligned} \mathscr{L} = {} & \mathbb{E}_q[\log p(x|z, \varphi)] + (\mathbb{E}_q[\log p(z|\theta_{dk})] - \mathbb{E}_q[\log q(z)]) \\ & + (\mathbb{E}_q[\log p(\theta_d|\varepsilon)] - \mathbb{E}_q[\log q(\theta_d|\tilde{\varepsilon})]) + (\mathbb{E}_q[\log p(\varphi_k|\zeta)] - \mathbb{E}_q[\log q(\varphi_k|\tilde{\zeta})]) \end{aligned}$$

$$\begin{aligned} \mathscr{L} = {} & \sum_{m=1}^{\mathscr{D}} \mathbb{E}_q \left[p(x_m|z_m, \varphi)\right] + \mathbb{E}_q \left[\log p(z_m|\theta_m)\right] \\ & - \mathbb{E}_q \left[\log q(z_m)\right] + \mathbb{E}_q \left[\log p(\theta_m|\varepsilon)\right] - \mathbb{E}_q \left[\log q(\theta_m)\right] + \left(\mathbb{E}_q \left[\log p(\varphi|\zeta)\right] - \mathbb{E}_q \left[\log q(\varphi)\right]\right) / \mathscr{D} \end{aligned} \tag{239}$$

The steps from (240) to (242) aim to summarize these expectations to facilitate the coordinate ascent framework when we compute the update equations as detailed in the Appendix.

$$\begin{aligned} & \mathbb{E}_q[\log p(z|\theta_{dk})] - \mathbb{E}_q[\log q(z)] \\ & \quad = \{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\mathbb{E}_q[\log \theta_{dk}] - \log \gamma_{nk}\right) + \sum_{n=1}^{N} \gamma_{n(K+1)} (\mathbb{E}_q[\log(\theta_{d(K+1)})] - \log \gamma_{n(K+1)})\} \end{aligned} \tag{240}$$

Table 6.1: Stochastic variational learning of LBLA [1]

```
General stochastic variational method
◇ Input:
        ◇ Corpus of $\mathscr{D}$ documents
        ◇ Step size $\rho_t = (\tau + t)^{-\kappa}$
        ◇ Maximum iteration $I$
        For $t = 1 : I$
         ◇ Select randomly a document $d_i$ from the corpus $\mathscr{D}$
         ◇ Compute noisy natural gradients including
          its local variational information $\Theta_{dt}$ using $\mathscr{D}$ copies
               $\hat{\nabla}_\Phi \ell(\Phi) = -\Phi + \mathscr{D}\Theta_{dt}$
         ◇ Update the global variational parameter (topic)

               $\Phi \leftarrow \Phi + \rho_t \hat{\nabla}_\Phi \ell(\Phi)$
◇ Output: $\Phi$
```

$$\begin{aligned}
&\mathbb{E}_q[\log p(\theta_d|\varepsilon)] - \mathbb{E}_q[\log q(\theta_d|\tilde{\varepsilon})] \\
&= \left[\sum_{k=1}^{K}(\alpha_k - \tilde{\alpha}_k)\mathbb{E}_q[\log \theta_{dk}]\right] + (\alpha - \tilde{\alpha})\mathbb{E}_q\left[\log\left(\sum_{k=1}^{K}\theta_{dk}\right)\right] + (\beta - \tilde{\beta})\mathbb{E}_q\left[\log\left(1 - \sum_{k=1}^{K}\theta_{dk}\right)\right] \\
&\quad + \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) + \log\Gamma(\alpha + \beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) - \log\Gamma\left(\sum_{k=1}^{K}\tilde{\alpha}_k\right) \\
&\quad - \log\Gamma(\tilde{\alpha} + \tilde{\beta}) + \log\Gamma(\tilde{\alpha}) + \log\Gamma(\tilde{\beta}) + \sum_{k=1}^{K}\log\Gamma(\tilde{\alpha}_k) \qquad (241)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_q[\log p(\varphi_k|\zeta)] - \mathbb{E}_q[\log q(\varphi_k|\tilde{\zeta})] \\
&= \left[\sum_{v=1}^{V}(\lambda_{kv} - \tilde{\lambda}_{kv})\mathbb{E}_q[\log \varphi_{kv}]\right] + (\lambda - \tilde{\lambda})\mathbb{E}_q\left[\log(\sum_{v=1}^{V}\varphi_{kv})\right] + (\eta - \tilde{\eta})\mathbb{E}_q\left[\log(1 - \sum_{v=1}^{V}\varphi_{kv})\right] \\
&\quad + \log\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right) + \log\Gamma(\lambda + \eta) - \log\Gamma(\lambda) - \log\Gamma(\eta) - \sum_{v=1}^{V}\log\Gamma(\lambda_v) \\
&\quad - \log\Gamma\left(\sum_{v=1}^{V}\tilde{\lambda}_{kv}\right) - \log\Gamma(\tilde{\lambda} + \tilde{\eta}) + \log\Gamma(\tilde{\lambda}) + \log\Gamma(\tilde{\eta}) + \sum_{v=1}^{V}\log\Gamma(\tilde{\lambda}_{kv}) \qquad (242)
\end{aligned}$$

6.3.4.1 Maximizing the smoothed LBLA's lower bound with respect to the variational parameters

This section computes the variational update equations. We place a BL prior on the global topic parameter in contrast to the method in [3, 35]. This leads to a smoothed LBLA

Table 6.2: Batch variational method

Complete Conditionals	Variational Updates
STEP 1: Complete conditionals distributions	
for exponential family	STEP 2: Compute expectations
$p(z_{dn} = k \vert \theta_d, \varphi, x_{dn})$ and $p(z_{dn} = K+1 \vert \theta_d, \varphi, x_{dn})$	$\gamma_{dn}^{k} = \mathbb{E}_q[z_{dn}^{k}]$ and γ_{dn}^{K+1}
using (244)	using (295)
$p(\theta_d \vert z_d) = BL(\bar{\alpha}_1, ..., \bar{\alpha}_K, \bar{\alpha}, \bar{\beta})$	$q(\theta_d \vert \tilde{\varepsilon}) = BL(\bar{\alpha}_1, ..., \bar{\alpha}_K, \bar{\alpha}, \bar{\beta})$
with hyperparameters using (246)	with hyperparameters using (245)
where $\theta_{d(K+1)} = 1 - \sum_k^K \theta_{dk}$	
with $\sum_k^K \theta_{dk} < 1$	
$p(\varphi_k \vert z = k, x) = BL(\bar{\lambda}_{k1}, ..., \bar{\lambda}_{kV}, \bar{\lambda}, \bar{\eta})$	$q(\varphi_k \vert \tilde{\zeta}) = BL(\tilde{\lambda}_{k1}, ..., \tilde{\lambda}_{kV}, \tilde{\lambda}, \tilde{\eta})$
with hyperparameters using (248)	with hyperparameters using (247)
where $\varphi_{k(V+1)} = 1 - \sum_{v=1}^{V} \varphi_{kv}$ with $\sum_{v=1}^{V} \varphi_{kv} < 1$	

model. Furthermore, the BL is asymmetric (non uniform base measure) to characterize heterogeneity in the topics. We are using the lower bound, the ELBO (evidence lower bound) to characterize coordinate ascent update equations. We derive the new update equations (both on the corpus and document level) for the parametric LBLA in this section as it will be useful when we apply our BNP prior using HDP-LBLA topic model. So this actually shows that this model is an extension to the work in [3, 35].

In the following section, we define the partial derivatives with respect to the model variational parameters (as we use BL priors both on the corpus and document parameters) and the latent variables z_{nk}. From these partials, we obtain the variational update equations. For instance, the coordinate ascent (ELBO) of γ_{nk} is defined as $\mathscr{L}(\gamma_{nk})$ while its partial derivative with respect to γ_{nk} is $\mathscr{L}'(\gamma_{nk}) = \frac{\partial \mathscr{L}}{\partial \gamma_{nk}}$. The corresponding partial derivative is set to zero from which we obtain the update equation for γ_{nk}. The corpus and documents variational update equations are computed and summarize in the Appendix.
Below is a list of some useful variational expectations for the variational update equations:
$\theta_{dk} \sim \mathrm{BL}(\alpha_k, ..., \alpha_K, \alpha, \beta)$
$\left(\sum_{k=1}^{K} \theta_{dk}\right) \sim \mathrm{Beta}(\alpha, \beta)$
$\mathbb{E}_q\left[\sum_{k=1}^{K} \theta_{dk}\right] = \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}}$

$$\mathbb{E}_q\left[(1-\sum_{k=1}^{K}\theta_{dk})\right]=\frac{\tilde{\beta}}{\tilde{\alpha}+\tilde{\beta}}$$
$$\mathbb{E}_q\left[\log(\sum_{k=1}^{K}\theta_{k})\right]=\Psi(\tilde{\alpha})-\Psi(\tilde{\alpha}+\tilde{\beta})$$
$$\mathbb{E}_q\left[\log(1-\sum_{l=1}^{K}\theta_{dk})\right]=\Psi(\tilde{\beta})-\Psi(\tilde{\alpha}+\tilde{\beta})$$
$$\mathbb{E}_q[\log\theta_{dk}]=\Psi(\tilde{\alpha})-\Psi(\tilde{\alpha}+\tilde{\beta})+\Psi(\tilde{\alpha}_k)-\Psi\left(\sum_{d=1}^{K}\tilde{\alpha}_k\right)\varphi_{kv}\sim \mathrm{BL}(\lambda_{k1},...,\lambda_{kV},\lambda,\eta)$$
$$\left(\sum_{v=1}^{V}\varphi_{kv}\right)\sim \mathrm{Beta}(\lambda,\eta)$$
$$\left(1-\sum_{v=1}^{V}\varphi_{kv}\right)\sim \mathrm{Beta}(\eta,\lambda)$$
$$\mathbb{E}_q\left[\sum_{v=1}^{V}\varphi_{kv}\right]=\frac{\tilde{\lambda}}{\tilde{\lambda}+\tilde{\eta}}$$
$$\mathbb{E}_q\left[1-\sum_{v=1}^{V}\varphi_{kv}\right]=\frac{\tilde{\eta}}{\tilde{\lambda}+\tilde{\eta}}$$
$$\mathbb{E}_q\left[\log\left(\sum_{k=1}^{K}\varphi_{kv}\right)\right]=\Psi(\tilde{\lambda})-\Psi(\tilde{\lambda}+\tilde{\eta})$$
$$\mathbb{E}_q\left[\log\left(1-\sum_{v=1}^{V}\varphi_{kv}\right)\right]=\Psi(\tilde{\eta})-\Psi(\tilde{\lambda}+\tilde{\eta})$$
$$\mathbb{E}_q[\log\varphi_{kv}]=\Psi(\tilde{\lambda})-\Psi(\tilde{\lambda}+\tilde{\eta})+\Psi(\tilde{\lambda}_{kv})-\Psi\left(\sum_{v=1}^{V}\tilde{\lambda}_k\right)$$

The topics models following the LDA architecture operate with three main conditional distributions (posteriors): $p(z_{dn}=k|\theta,\varphi_k,x)$, $p(\theta_d|z_d)$, and $p(\varphi_k|z,x)$.

The work in [1] shows that for exponential family distributions, variational parameters could be obtained by taking expectations of the natural parameters of their corresponding complete conditionals. We apply this method for our proposed asymmetric and smoothed LBLA model, the alternative to the (symmetric) LDA topic model. It will be also implemented in our HDP-LBLA model when we evaluate infinite dimensional complete conditionals.

First, we estimate the complete conditionals for the parametric LBLA by following the conditional of LDA model. In LDA, we have:

$$\begin{aligned} p(z=k|\theta_d,\varphi,x) &\propto p(x|z=k,\varphi)p(z=k|\theta_d)p(\theta)p(\varphi) \\ &\propto p(x|z=k,\varphi)p(z=k|\theta_d) \\ &\propto (\theta_{dk}\varphi_{kv}) \\ &\propto \exp\{\log(\theta_{dk}\varphi_{kv})\} \\ &\propto \exp\{\log\theta_{dk}+\log\varphi_{kv}\} \end{aligned} \tag{243}$$

The observation here is that the complete conditional of topic assignment only groups the parameters that have common z_{dn} as indicator (in $p(x|z,\varphi)$ and $p(z|\theta)$) or common γ_{dn} in variational setting between $\mathbb{E}_q[\log p(x|z,\varphi)]$ and $\mathbb{E}_q[\log p(z|\theta)]$. Evaluating the complete conditionals of the latent variables for LBLA is very complex due to the composition between the Beta distribution and the Liouville of the second kind distribution that ultimately form the BL prior [11]. The form is not as straightforward as in the LDA. From LDA, using Dirichlet distribution, we know that coordinate ascent variational inferences iterate between updating the local context variational parameters (the local per-document topic proportions Dirichlet parameters and the per-word topic assignment multinomial parameters) and updating the global variational parameters.

Importantly, these latent update equations are obtained from expectations of the natural parameters of the complete conditionals. This relationship between complete conditionals natural parameters and variational parameters allows us to recover complex complete conditionals such as the one for the topic assignment for our LBLA variational framework.

Using the variational multinomial parameter update from (296), we can simply express the corresponding complete conditionals for topic assignments as:

$$\begin{cases} p(z_{dn} = k|\theta, \varphi, x_{dn}) \propto \exp\{\log \theta_{dk} + \log \varphi_{kv} + \log(\varphi_{k(V+1)})\} \\ p(z_{dn} = K+1|\theta, \varphi, x_{dn}) \propto \exp\{\log(\theta_{d(K+1)})\} \end{cases} \tag{244}$$

Using our variational updates, we can deduct updates for the complete conditionals as shown in Table 6.2 for LBLA. The variational updates are summarized in Table 6.1. We can see the difference between LDA method in [1] and the asymmetric and smoothed LBLA just by examining the updates in the complete conditionals (posteriors) and variational posteriors. The posterior $p(\theta_d|z_d)$ is $\mathrm{BL}(\bar{\alpha}_{d1}, ..., \bar{\alpha}_{dK}, \bar{\alpha}, \bar{\beta})$ while $p(\varphi_k|z, x)$ is $\mathrm{BL}(\bar{\lambda}_{k1}, ..., \bar{\lambda}_{vK}, \bar{\lambda}, \bar{\eta})$. The document variational updates equations in our work are:

$$\tilde{\alpha}_{dk} = \alpha_{dk} + \sum_{n=1}^{N} \gamma_{dn}^{k} \qquad \tilde{\alpha} = \alpha + \sum_{n=1}^{N}\sum_{k=1}^{\mathscr{D}} \gamma_{dn} \qquad \tilde{\beta} = \beta + \sum_{n=1}^{N}\sum_{d=1}^{\mathscr{D}} \gamma_{dn}^{K+1} \tag{245}$$

The complete conditional parameters updates are the following:

$$\bar{\alpha}_{dk} = \alpha_{dk} + \sum_{n=1}^{N} z_{dn}^{k} \qquad \bar{\alpha} = \alpha + \sum_{n=1}^{N}\sum_{k=1}^{\mathscr{D}} z_{dn} \qquad \bar{\beta} = \beta + \sum_{n=1}^{N}\sum_{d=1}^{\mathscr{D}} z_{dn}^{K+1} \tag{246}$$

Similarly, in the word-topic distribution, the corpus variational parameters are in (247) while (248) shows the complete conditionals of the topics (global parameters).

$$\tilde{\lambda}_{kv} = \lambda_{kv} + \sum_{n=1}^{N} \gamma_{dn}^{k} x_{dn}^{v} \qquad \tilde{\lambda} = \lambda + \sum_{n=1}^{N}\sum_{k=1}^{\mathscr{D}} \gamma_{dn} x_{dn} \qquad \tilde{\eta} = \eta + \sum_{n=1}^{N}\sum_{d=1}^{\mathscr{D}} \gamma_{dn} x_{dn}^{V+1} \tag{247}$$

$$\bar{\lambda}_{kv} = \lambda_{kv} + \sum_{n=1}^{N} z_{dn}^{k} x_{dn}^{v} \qquad \bar{\lambda} = \lambda + \sum_{n=1}^{N}\sum_{k=1}^{\mathscr{D}} z_{dn} x_{dn} \qquad \bar{\eta} = \eta + \sum_{n=1}^{N}\sum_{d=1}^{\mathscr{D}} z_{dn} x_{dn}^{V+1} \tag{248}$$

The variational parameters are expectations of the natural parameters of their corresponding complete conditionals. This connection between complete conditionals and variational parameters will allow easy implementation of HDP-LBLA as we move from finite parametric topic model to nonparametric (infinite dimensional) topic model as we will show. The update equations from (293) to (302) obtained are then rewritten using indicator z_{dn}^{k} and its variational $\mathbb{E}_q[z_{dn}^{k}] = \gamma_{dn}^{k}$ instead of z_{nk} and its $\mathbb{E}_q[z_{nk}] = \gamma_{nk}$ to characterize our stochastic framework at document level as shown in Table 6.2.

6.3.4.2 Stochastic optimization and convergence framework of the LBLA

We implement a stochastic optimization using the natural gradient of the ELBO, and we study the convergence of our asymmetric smoothed LBLA. We first derive the appropriate ELBO for the the stochastic optimization in (249) which allows easy implementation of the natural gradient method. We use the objective function $\mathscr{L}(\tilde{\varepsilon}, \gamma, \tilde{\zeta}) = \sum_d \mathscr{L}_d(n_d, (\tilde{\alpha}_d, \tilde{\alpha}, \tilde{\beta})_d, \gamma_d, (\tilde{\lambda}_k, \tilde{\lambda}, \tilde{\eta})_k)$ where

$$\begin{aligned} \mathscr{L}_d = \mathbb{E}_q\left[p(x_d|z_d, \varphi)\right] + \mathbb{E}_q\left[\log p(z_d|\theta_d)\right] - \mathbb{E}_q\left[\log q(z_d)\right] \\ + \mathbb{E}_q\left[\log p(\theta_d|\varepsilon)\right] - \mathbb{E}_q\left[\log q(\theta_d)\right] + \left(\mathbb{E}_q\left[\log p(\varphi|\zeta)\right] - \mathbb{E}_q\left[\log q(\varphi)\right]\right)/\mathscr{D} \end{aligned}$$

so that it leads to:

$$\mathscr{L} = \sum_{d=1}^{\mathscr{D}} \mathscr{L}_d = \mathscr{D}\mathbb{E}_d[\mathscr{L}_d] = \mathbb{E}_d[\mathscr{D}\mathscr{L}_d] \tag{249}$$

When summing over the documents using [167, 65], we identified the per-corpus terms (the corpus wide-terms) and divided them by $\mathscr{D}$ (the total number of documents). We therefore show that our lower bound could also be expressed as an expectation over the distribution of the data (empirical) which characterizes a variational lower bound computed using $\mathscr{D}$ copies of a document d.

The lower bound is also expressed as a function of variational parameters including the document count variables. To find a maximum of our objective function we apply a step size ρ_t in the direction of the natural gradient to speed up process. The natural gradient also characterizes the information geometry of the parameter space which uses the Riemannian metric or Fisher information matrix to guide the standard gradient. The metric locally uses the KL divergence between distributions. We will provide more information on the natural gradient in section 6.4 when we implement for the stochastic HDP-LBLA. Following the work in [65] about the gradient method for optimizing the global topic parameters λ, for instance, we also proved that premultiplying that gradient $\frac{\partial \mathscr{L}(n_d,(\tilde{\alpha}_d,\tilde{\alpha},\tilde{\beta})_d,\psi_d,(\tilde{\lambda}_k,\tilde{\lambda},\tilde{\eta})_k)}{\partial \tilde{\lambda}_k}$ by the inverse of the Fisher information matrix $\left(-\frac{\partial^2 \log \varphi_{kv}}{\partial \tilde{\lambda}_k \tilde{\lambda}_k^T}\right)^{-1}$ leads to a stochastic method which implements a noisy natural gradient of the proposed ELBO using $\mathscr{D}$ copies of document d along with ρ_t as step size (learning rate) as shown below.
$\mathscr{L} = \mathscr{L}(n_d, (\tilde{\alpha}_d, \tilde{\alpha}, \tilde{\beta})_d, \psi_d, (\tilde{\lambda}_k, \tilde{\lambda}, \tilde{\eta})_k)$

$$\rho_t\mathscr{D}\left[\left(-\frac{\partial^2 \log \varphi_k}{\partial \tilde{\lambda}_k \tilde{\lambda}_k^T}\right)^{-1} \frac{\partial \mathscr{L}}{\partial \tilde{\lambda}_k}\right]_v = \rho_t\mathscr{D}\left(\frac{-\tilde{\lambda}_k}{\mathscr{D}} + \frac{\lambda_k}{\mathscr{D}} + n_{tv}\gamma_{tvk}\right) \tag{250}$$

$$= \rho_t(-\tilde{\lambda}_{kv} + \lambda_{kv} + \mathscr{D}n_{tv}\gamma_{tvk}) \tag{251}$$

$$= \rho_t\hat{\nabla}_\lambda \ell(\lambda) \tag{252}$$

$\hat{\nabla}_\lambda \ell(\lambda)$ includes the inverse of Fisher information or the inverse of the Riemannian metric. From (252), we can also expand it when we add $\tilde{\lambda}_{kv}$ to generate the well known online average approach as follow:

$$\begin{aligned}\rho_t\hat{\nabla}_\lambda \ell(\lambda) + \tilde{\lambda}_{kv} &= \tilde{\lambda}_{kv} - \rho_t\tilde{\lambda}_{kv} + \rho_t\lambda_{kv} + \rho_t\mathscr{D}n_{tv}\gamma_{tvk} \\ &= (1-\rho_t)\tilde{\lambda}_{kv} + \hat{\tilde{\lambda}}_{kv}\end{aligned}$$

We can deduct the stochastic updates using noisy natural gradient method with $\mathscr{D}$ copies for the global variational parameters as:

$$\begin{cases}\tilde{\lambda}_{kv} \leftarrow \tilde{\lambda}_{kv} + \rho_t\hat{\nabla}_{\tilde{\lambda}}\ell(\tilde{\lambda}) \\ \tilde{\lambda} \leftarrow \tilde{\lambda} + \rho_t\hat{\nabla}_\lambda \ell(\lambda) \\ \tilde{\eta} \leftarrow \tilde{\eta} + \rho_t\hat{\nabla}_\eta \ell(\eta)\end{cases} \tag{253}$$

where

$$\begin{cases}\hat{\nabla}_{\tilde{\lambda}_{kv}}\ell(\tilde{\lambda}_{kv}) = -\tilde{\lambda}_{k_{v=w}} + \lambda_{kv} + \mathscr{D}\sum_{n=1}^N \gamma_{nd}^k x_{dn} \\ \hat{\nabla}_{\tilde{\lambda}}\ell(\tilde{\lambda}) = -\tilde{\lambda} + \lambda + \mathscr{D}\sum_{n=1}^N \gamma_{dn} \\ \hat{\nabla}_{\tilde{\eta}}\ell(\tilde{\eta}) = -\tilde{\eta} + \eta + \mathscr{D}\sum_{n=1}^N \gamma_{dn} x_{dn}^{(V+1)}\end{cases} \tag{254}$$

In online framework (Table 6.3), given the initial values of the global variational parameters, we randomly select a document j which is used to compute the local parameters that include the variational responsibility distribution and the document parameters. The local context of the document is therefore used to compute the natural gradient of the global parameters (corpus parameters) from the variational ELBO using $\mathscr{D}$ copies of document j. The global variables in stochastic optimization update method follow their natural gradients with a step size $\rho_t \triangleq (\tau + t)^{-\kappa}$ where for convergence $\kappa \in (0.5 \;\; 1]$. The variable κ monitors the rate at which old values are forgotten while τ slows down the early iterations t of the algorithm [1, 65, 167, 107]. Since $\sum_{t=0}^{\infty} \rho_t = \infty$ and $\sum_{t=0}^{\infty} \rho_t^2 < \infty$ therefore, $\tilde{\lambda}_k$, $\tilde{\lambda}$ and $\tilde{\eta}$ each converges to a stationary point with each of their respective gradients $(\hat{\nabla}_{\tilde{\lambda}_{kv}} \ell(\tilde{\lambda}_{kv}), \hat{\nabla}_{\tilde{\lambda}} \ell(\tilde{\lambda}), \hat{\nabla}_{\tilde{\eta}} \ell(\tilde{\eta}))$ converges to zero. As stochastic variational framework at document level, only the global parameters (here the global topics) are updated [1]. In this stochastic optimization method, we could set a minibatch technique to reduce noise by using multiple samples (documents) at a time. Instead of computing the natural gradient of $\mathscr{D}\mathscr{L}_d$, we could implement the natural gradient of a minibatch defined as:

Table 6.3: Stochastic Variational Inference for parametric LBLA topic model

```
Stochastic Variational Method for Asymmetric LBLA

        ◇ INITIALIZATIONS:
◇ Choose a number of topics $K$
◇ Set $\rho_t$ such that $\rho_t = (\tau + t)^{-\kappa}, t = 1, t \leftarrow t + 1$.
◇ Initialize (corpus) global variational parameters $\tilde{\lambda}_k, \tilde{\lambda}, \tilde{\eta}$
◇ Draw a document $d$ uniformly from the corpus
◇ Initialize the document local variational parameters $\tilde{\alpha}_{dk}, \tilde{\alpha}_d, \tilde{\beta}_d$
        ◇ E-STEP: Evaluate the local context of a document
◇ Update $\gamma_{dn}^{k}$ and $\gamma_{dn}^{K+1}$ using (293) and (294)
◇ Document variational parameters update $(\tilde{\alpha}_{dk}, \tilde{\alpha}_d, \tilde{\beta}_d)$ using (245)
        ◇ M-STEP:
◇ Compute natural gradients using $\mathscr{D}$ copies of document $d$ from (254)
◇ Global topic variational paramaters update $(\tilde{\lambda}_k, \tilde{\lambda}, \tilde{\eta})$ using (253)
◇ Until convergence
◇ Output: $\tilde{\lambda}_k, \tilde{\lambda}, \tilde{\eta}$
```

$$\mathscr{L}_s = \frac{\mathscr{D}}{S} \sum_{d \in S} \mathscr{L}_d \tag{255}$$

where S is the subset of documents and S the cardinality of S (the number of documents in S). We can see that when $S = 1$, we have a stochastic method of size one (one document at a time). When $S > 1$, we have a regular minibatch framework (using multiple documents at a time). When $S = \mathscr{D}$ with the forgetting rate $\kappa = 0$, then we have a batch-based stochastic variational scheme. This shows the flexibility of the minibatch method for characterizing online stochastic variational LBLA.

6.3.4.3 Predictive models

To compute the per-word log predictive probability $\mathcal{L}$ in (287), we estimate the expectations in (286) as we marginalize out the models parameters. We compute the expectation of the documents BL distributions $\mathbb{E}_q[\theta_{dk}|\tilde{\varepsilon}]$ using the held out data (previoulsy unseen text documents). However, we estimate $\mathbb{E}_q\left[\varphi_k|\tilde{\zeta}\right]$ from the training set $\mathscr{D}_{tr}$ which is then maintained fixed during prediction. It represents the expectation of the corpus BL distribution. Given $\tilde{\varepsilon} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_K, \tilde{\alpha}, \tilde{\beta})$, we evaluate the predictive estimate $\mathbb{E}_q[\theta_k|\tilde{\varepsilon}]$ using:

$$\mathbb{E}_q[\theta_{dk}|\tilde{\varepsilon}] = \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} \frac{\tilde{\alpha_k}}{\sum_{k=1}^{K} \alpha_k} \tag{256}$$

where (245) allows us to express the documents BL's parameters $\tilde{\alpha}_k$, $\tilde{\alpha}$, and $\tilde{\beta}$. As we can notice in (256), marginalizing over the parameters ultimately shows the clustering property as it introduces distributions over partitions, which is reminiscent of the multivariate Polya urn process [157]. Here, we have a more general version compared to the one from LDA with Dirichlet prior [1].

6.4 HDP-LBLA model

In this proposed approach, we place a hierarchical Bayesian nonparametric prior on the documents multinomials. Documents in topic models are naturally defined as mixtures over topics, and the objective is to provide a prior that not only solves the problem of optimal number of topics (model selection) but also allows documents to share the global topics. It is understood that each document exhibits topics in different proportions. The possibility of sharing global topics therefore connects all corpus documents to the global probability measure (that provides the set of all topics). In hierarchical mixture modeling, this setting automatically makes the HDP the right Bayesian nonparametric prior to model efficiently the topic mixtures.

Our proposed method is mainly an alternative to the standard and parametric LDA which is very restrictive both in model selection and sharing ability of topics due to the limitations [3, 5, 8] of its Dir prior. Among these limitations, the symmetric Dir does not provide variability and heterogeneity of topics. This is because under its symmetric LDA, topics are more likely to exhibit same frequency. Therefore, symmetric LDA could not be applied in a real life scenario when, for instance, information of the most relevant topics are needed for a better compression algorithm. To encourage variability, heterogeneity where coarser topics can combine with finely grained and detailed ones, we propose asymmetric BL distribution as a global measure and as an alternative to the symmetric Dir prior in LDA. The setting enhances the asymmetric nature of the GEM priors in our HDP-LBLA. Importantly, such flexibility in the topic structure offers the possibility of a fast detection of most relevant topics and as a result provides an alternative to model selection. We propose a stochastic variational approach, and it summarizes the model to SV-HDP-LBLA (stochastic variational Bayes using HDP-LBLA). In implementing the SV-HDP-LBLA, we first constructed the finite dimensional SV-LBLA and then we simply accommodate it in infinite dimensional space where the multinomials are modeled by our nonparametric prior, the BL-based HDP. We characterize the Sethuraman's stick breaking method on both levels of the HDP-LBLA model. The proposed stochastic optimization with minibatches uses noisy estimates of the natural gradient of our ELBO (objective function).

6.4.1 Generative process for two level-HDP using asymmetric BL

In our BNP topic model, the BL-based HDP replaces the standard Dirichlet prior in LDA as it models documents multinomials. The asymmetric BL acts as a diffuse base measure H and provides the global topics. We use the Sethuraman's stick breaking construction for the two level hierarchical BL-based HDP topic model in this work. We implemented two stick-breaking constructions, one for each level (corpus and document level). Similar to the LDA case in [1], the generative process for our HDP with BL as a top-level base measure is given as follows:
Let $\zeta = (\lambda_{k1}, ..., \lambda_{kV}, \lambda, \eta)$
Draw an infinite number of topics $\varphi_k \sim BL(\zeta)$ for $k \in \{1, 2, 3, ...\}$
Draw corpus breaking proportions $\varsigma_k^{'} \sim Beta(1, \hbar)$ for $k \in \{1, 2, 3, ...\}$
For each document d
$a-$Draw a document-level topic indices $\xi_{di} \sim Mult(\varsigma)$ for $i \in \{1, 2, 3, ...\}$
$b-$Draw the document breaking proprtions $\phi_{di}^{'} \sim Beta(1, \varrho)$ for $i \in \{1, 2, 3, ...\}$
$c-$For each word x_{dn}
$i)$ Draw topic assignment $z_{dn} \sim Mult(\phi_d)$
$ii)$ Draw a word $x_{dn}|\xi_{di}, z_{dn}, \varphi \sim Mult(\varphi_{\xi d,_{z_{dn}}})$

where $Mult$ is the multinomial distribution. This generative process naturally defines the batch HDP-LBLA.

6.4.2 Sethuraman's stick-breaking method for HDP-LBLA

A hierarchical Dirichlet process is a distribution over a collection of probability measures over a measurable space $(\Theta, \mathscr{B}(\Theta))$. It characterizes the connection between documents as they share global clusters (topics), but at different proportions. In this chapter, our two level hierarchical Dirichlet process is similar to the work in [14], [193], [1], [167]. The difference is that we use BL prior instead of Dir. We define the global probability measure G_0:

$$G_0|\hbar, H \sim DP(\hbar, H) \tag{257}$$

For each document d, we draw with probability one:

$$G_d|\varrho, G_0 \sim DP(\varrho, G_0) \tag{258}$$

The base measure at the corpus level (top-level) DP is a fully asymmetric BL distribution whose atoms are the topics (global variables). From the first DP, we draw G_0 with probability one and then use it as a random global (base) measure in the second level (document-level) DP from which we draw G_d random probability measures, each with probabilty one. They are conditionally independent given the random global measure G_0. First, a draw G_0 at corpus-level DP$(\hbar, H)$ following a stick-breaking scheme can be characterized as follows:

$$\varsigma_k^{'} \sim Beta(1,\hbar) \qquad \varsigma_k = \varsigma_k^{'} \prod_{l=1}^{k-1}(1-\varsigma_l^{'}) \qquad \varphi_k \sim H \qquad G_0 = \sum_{k=1}^{\infty} \varsigma_k \delta_{\varphi_k} \tag{259}$$

where H is the asymmetric BL distribution, φ_k are the global topics such that φ_k are drawn from the base measure H; δ_φ is a probability measure concentrated at φ. The sequence ς is a random probability measure on the positive integers [14, 15]. In other words G_0 has weights or probability masses $\varsigma = \{\varsigma_k\}_{k=1}^{\infty}$ such that $\sum_{k=1}^{\infty} \varsigma_k = 1$. The support of the discrete distribution G_0 is the set of atoms $\varphi = \{\varphi_k\}_{k=1}^{\infty}$. Furthermore, as ς satisfies (259), we have $\varsigma \sim GEM(\hbar)$ where GEM stands for Griffiths, Engen and McCloskey [183]. They characterize the GEM distributions of the random weights ς. From the document-level DP conditioned on G_0, we constructed the local random measure G_d as follows:

$$\phi_{dk}^{'} \sim Beta(1,\varrho) \qquad \phi_{dk} = \phi_{dk}^{'} \prod_{j=1}^{k-1}(1-\phi_{dj}^{'}) \qquad G_d = \sum_{k=1}^{\infty} \phi_{dk} \delta_{\varphi_k} \tag{260}$$

This stick-breaking construction allows both G_d and G_0 distributions to share the same support (topics). However, sharing directly the same atoms makes their weights tightly coupled, and it is a situation that negatively affects variational inference and coordinate ascent methods as it makes them difficult and challenging as shown in [194], [16], [167] where closed-form variational updates are not possible. A solution has been to draw the document-level topics (atoms) from G_0 instead while still maintaining $\phi_d \sim GEM(\varrho)$. Now, we have $\Omega_{dt} \sim G_0$ so that $G_d = \sum_{t=1}^{\infty} \phi_{dt} \delta_{\Omega_{dt}}$.
At the document level, we finally get:

$$\Omega_{dt} \sim G_0 \qquad \phi_{dt}^{'} \sim Beta(1,\varrho) \qquad \phi_{dt} = \phi_{dt}^{'} \prod_{j=1}^{t-1}(1-\phi_{dj}^{'}) \qquad G_d = \sum_{l=1}^{\infty} \phi_{dt} \delta_{\Omega_{dt}} \tag{261}$$

such that $\Omega_{dt} = (\Omega_{dt})_{t=1}^{\infty}$. However, we need indicators to connect to the global atoms. This is given by $\xi_{dt} = (\xi_{dt})_{t=1}^{\infty}$ where $\xi_{dt} \sim Mult(\varsigma)$. The variables ξ_{dt} as indicators connect the document-level to the corpus level. They index the corpus level topics that correspond to Ω_{dt} such that $\Omega_{dt} = \varphi_{\xi_{dt}}$. We can therefore see with $\Omega_{dt} = \varphi_{\xi_{dt}}$ that the global topics at the corpus level are still shared at the document level. To generate a word x_{dn} in a document d in our BL-based HDP model, we first draw $\xi_{dt} \sim Mult(\varsigma)$; then, given ξ_{dt}, we draw the topic $\Omega_{dt} = \varphi_{\xi_{dt}}$ from G_0. Then, we sample its topic assignment $z_{dn} \sim Mult(\phi_d)$ (using the document topic proportions or weights ϕ_d). Then, the word x_{dn} is drawn from the top level topic space indexed by indicator ξ such that $x_{dn}|z_{dn}, \varphi_{\xi_{dt}} \sim Mult(\varphi_{\xi_{d z_{dn}}})$ [167], [182], [1]. In this representation, φ is drawn from our BL distribution.

6.4.3 Inference for HDP-LBLA

The HDP-LBLA means that inference is governed by the LBLA topic model using the HDP prior (nonparametric) in infinite dimensional space as it models infinite dimensional topic mixtures for each document. The most important task in Bayesian inference for parametric and nonparametric models is the computation of posterior distributions. where we estimate the posterior distribution over the number of topics that describes the observed data. In general, in a latent model as in our case, the number of topics is unknown in advance.

In nonparametric setting, the exact posterior distribution in HDP (in infinite dimensional space) is expected to be intractable. We implement a variational framework with two levels of trunctations (at the corpus level and at the document level). This is because BNP models contain infinite number of hidden variables, so they cannot be fully estimated by variational distributions as this will be equivalent to optimizing over infinite number of variational free parameters. Though, one of the advantages of variational methods is that the variational distributions are in the same family as the complete conditional distributions (which are exponential family and therefore have closed-form solutions).

In this Bayesian nonparametric model using HDP, our hidden variables include the top (corpus)-level stick-breaking proportions ς, the document-level stick-breaking proportions $\phi^{'}_{di}$, the latent indictator variables ξ_{di} for each G_d, the atom/topic distributions φ_k, and topic index z_{dn} for each word x_{dn}. For a stochastic framework, it is important to identify the global and local variables: the global variables are the topics φ_k including the corpus level breaking proportions $\varsigma^{'}_k$. Local variables are the document-level topic indices ξ_{di} and breaking proportions $\phi^{'}_{di}$ including the latent variables z_{dn} and the words x_{dn}. We implement a fully factorized (mean-field) variational Bayes inference. Because of (294), we have to set the truncations at K and T. Furthermore, in general, in practice, the value T can be set smaller than K because each document requires a much lower number of topics than the case that considers the entire corpus which has a much larger set of topics. Therefore, $K >> T$.

Table 6.4: Complete conditionals

Complete conditionals for HDP-LBLA
Local topic assignment:
$p(z^i_{dn} = 1 \vert \phi^{'}_d, \varphi_{1:K}, x_{dn}, \xi_d)$ using (270)
Connector between global and local topics:
$p(\xi^k_d = 1 \vert \varsigma^{'}, \varphi_{1:K}, x_d, z_d)$ using (269)
Topic posterior distributions:
$p(\varphi_k \vert z, \xi, x) = BL(\bar{\lambda}_{k1}, ..., \bar{\lambda}_{kV}, \bar{\lambda}, \bar{\eta})$ using (267)
Conditionals on stick-breaking proportions:
$p(\varsigma^{'}_k \vert \xi) = Beta(\Delta_1, \Delta_2)$ using (268) (corpus breaking proportion)
$p(\phi^{'}_{di} \vert z_d) = Beta(\Upsilon_1, \Upsilon_2)$ using (271) (document breaking proportion)

Given our HDP-LBLA topic model's hidden variables, the variational distribution is expressed as $q(\varphi, \xi, \varsigma^{'}, z, \phi^{'}) = q(\varphi)q(\xi)q(\varsigma^{'})q(z)q(\phi^{'})$ with each factor extended as follows:

$$q(z) = \prod_{d=1}^{\mathscr{D}} \prod_{n=1}^{N} q(z_{dn} \vert \gamma_{dn}) \qquad q(\varphi) = \prod_{k=1}^{K} q(\varphi_k \vert \tilde{\zeta}) \qquad q(\xi) = \prod_{d=1}^{\mathscr{D}} \prod_{i=1}^{T} q(\xi_{di} \vert \mho_{di})$$

where the variational parameters $\mho_{di}$ and γ_{dn} are multinomials while $\tilde{\zeta}$ is a BL parameter.

$$q(\varsigma^{'}) = \prod_{k=1}^{K} q(\varsigma_k^{'}|\iota_k, b_k) \qquad\qquad q(\phi^{'}) = \prod_{d=1}^{\mathscr{D}} \prod_{i=1}^{T} q(\phi_{di}^{'}|\vartheta_{di}, r_{di})$$

where (ι_k, b_k) and (ϑ_{di}, r_{di}) are variational parameters of Beta distributions.

$$q(\varphi, \xi, \varsigma^{'}, z, \phi^{'}) = \prod_{k=1}^{K} q(\varphi_k|\tilde{\zeta})q(\varsigma_k^{'}|\iota_k, b_k) \prod_{d=1}^{\mathscr{D}} \prod_{i=1}^{T} q(\xi_{di}|\mho_{di})q(\phi_{di}^{'}|\vartheta_{di}, r_{di}) \prod_{n=1}^{N} q(z_{dn}|\gamma_{dn}) \quad (262)$$

Because variational posterior distributions approximate their corresponding complete conditional distributions (posterior) in mean-field (batch) variational inference, the free parameters of the variational distributions are usually obtained as expectations of the natural parameters of their corresponding complete conditionals specifically for exponential family distributions [1] as in our case.

For a collection of $\mathscr{D}$ documents, using the Jensen's inequality, we can compute the log marginal likelihood with the truncated variationals:

$$\log p(\mathscr{D}|\hbar, \varrho, \zeta) = \log \int_{\varphi, \varsigma^{'}, \phi^{'}} \sum_{\xi, z} p(\mathscr{D}, \varphi, \varsigma^{'}, \phi^{'}, \xi, z, x|\hbar, \varrho, \zeta) \quad (263)$$

$$\log p(\mathscr{D}|\hbar, \varrho, \zeta) \geq \mathbb{E}_q\left[\log p(\mathscr{D}, \varphi, \varsigma^{'}, \phi^{'}, \xi, z, x|\hbar, \varrho, \zeta)\right] - \mathbb{E}_q\left[\log q(\varphi, \xi, \varsigma^{'}, z, \phi^{'})\right] = \mathscr{F}(q) \quad (264)$$

The ELBO is obtained as:

$$\mathscr{F}(q) = \sum_j \mathbb{E}_q[\log p(x_j|\xi_j, z_j, \varphi)p(\xi_j|\varsigma^{'})p(z_j|\phi_j^{'})p(\phi_j^{'}|\varrho)p(\varphi|\zeta)p(\varsigma^{'}|\hbar)]$$
$$- \mathscr{H}(q(\xi_j)) - \mathscr{H}(q(z_j)) - \mathscr{H}(q(\phi_j^{'})) - \mathscr{H}(q(\varsigma_k^{'})) - \mathscr{H}(q(\varphi)) \quad (265)$$

This is also equivalent to:

$$\mathscr{F}(q) = \sum_j \mathbb{E}_q[\log p(x_j|\xi_j, z_j, \varphi)p(\xi_j|\varsigma^{'})p(z_j|\phi_j^{'})] - \mathscr{H}(q(\xi_j)) - \mathscr{H}(q(z_j))$$
$$- \mathscr{H}(q(\phi_j^{'})) + \mathbb{E}\left[\log p(\phi_j^{'}|\varrho)\right] + \mathbb{E}_q\left[\log p(\varphi|\zeta)\right] + \mathbb{E}_q\left[\log p(\varsigma^{'}|\hbar)\right] - \mathscr{H}(q(\varsigma_k^{'})) - \mathscr{H}(q(\varphi)) \quad (266)$$

with $\mathscr{H}(.)$ defining the entropy of the variational distribution.

6.4.3.1 Infinite dimensional complete conditionals, truncations, variational expectations, and updates

In the following, we express the infinite dimensional complete conditionals of the global topics/atoms φ_k, the corpus level stick-breaking proportions $\varsigma_k^{'}$, the indicator variables ξ_{di}, the latent variables z_{dn}, and document level stick-breaking proportions $\phi_d^{'}$ as shown from (267) to (271). The LBLA-based complete conditionals became infinite dimensional along with its natural parameters due to the HDP prior. These infinite dimensional complete conditionals provide the reason our proposed HDP-LBLA-based variational framework

requires truncations.
$p(\varphi_k|z, \xi, x) = BL(\bar{\zeta})$ such that:

$$\bar{\lambda}_{kv} = \lambda_{kv} + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{\infty} \xi_{di}^{k} \sum_{n=1}^{N} z_{dn}^{i} x_{n}^{v} \quad \bar{\lambda} = \lambda + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{\infty} \xi_{di}^{k} \sum_{n=1}^{N} z_{dn} x_{dn} \quad \bar{\eta} = \eta + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{\infty} \xi_{di}^{k} \sum_{n=1}^{N} z_{dn} x_{dn}^{V+1} \tag{267}$$

$$p(\varsigma_k^{'}|\xi) = Beta\left(1 + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{\infty} \xi_{di}^{k}, \hbar + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{\infty} \sum_{j>k} \xi_{di}^{j}\right) \tag{268}$$

$$p(\xi_d^i = 1|\varsigma^{'}, \varphi_{1:K}, x_d, z_d) \propto \exp\left\{ \log \varsigma_k + \sum_{n=1}^{N} z_{dn}^{i} \left(\log \varphi_{kv} + (1 - \log \left(\sum_{v}^{V} \varphi_{kv} \right) \right) \right\} \tag{269}$$

$$p(z_{dn}^i = 1|\phi_d^{'}, \varphi_{1:K}, x_{dn}, \xi_d) \propto \exp\left\{ \log \phi_{di} + \sum_{k=1}^{\infty} \xi_{di}^{k} \left(\log \varphi_{kv} + \log \left(1 - \sum_{v=1}^{V} \varphi_{kv} \right) \right) \right\} \tag{270}$$

$$p(\phi_d^{'}|z_d) = Beta\left(1 + \sum_{n=1}^{N} z_{dn}^{i}, \varrho + \sum_{n=1}^{N} \sum_{j>i} z_{dn}^{j}\right) \tag{271}$$

As shown above from (267) to (271), because the natural parameters of these complete conditionals are infinite dimensional, variational Bayesian inference with a tractable ELBO will be difficult to implement without truncations.

After the proposed truncations from (262) and using our complete conditional distributions, we express the variational parameters through expectations of the natural parameters of these complete conditionals for our exponential family based HDP-LBLA topic model (Table 6.4). Therefore, from (267), the corpus-level variational parameters of $q(\varphi_k|\tilde{\zeta})$ from the truncations become:

$$\begin{cases} \tilde{\lambda}_{kv} = \lambda_{kv} + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{T} \mathbb{E}_q[\xi_{di}^{k}] \sum_{n=1}^{N} \mathbb{E}_q[z_{dn}^{i}] x_{n}^{v} \\ \tilde{\lambda} = \lambda + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{T} \mathbb{E}_q[\xi_{di}^{k}] \sum_{n=1}^{N} \mathbb{E}_q[z_{dn}] x_{dn} \\ \tilde{\eta} = \eta + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{T} \mathbb{E}_q[\xi_{di}^{k}] \sum_{n=1}^{N} \mathbb{E}_q[z_{dn}] x_{dn}^{V+1} \end{cases} \tag{272}$$

Using the complete conditional $p(\varsigma_k^{'}|\xi)$ in (268), its corresponding variational distribution $q(\varsigma_k^{'}|\iota_k, b_k)$ has its Beta variational parameters truncated as:

$$\iota_k = 1 + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{T} \mathbb{E}_q[\xi_{di}^{k}] \qquad\qquad b_k = \hbar + \sum_{d=1}^{\mathscr{D}} \sum_{i=1}^{T} \sum_{j=k+1}^{K} \mathbb{E}_q[\xi_{di}^{j}] \tag{273}$$

Similarly, with $p(\phi_d^{'}|z_d)$ as in (271), its corresponding document's variational posterior distribution $q(\phi_d^{'}|\vartheta_{di}, r_{di})$ also has its variational Beta parameters expressed as:

$$\vartheta_{di} = 1 + \sum_{n=1}^{N} \mathbb{E}_q[z_{dn}^{i}] \qquad\qquad r_{di} = \varrho + \sum_{n=1}^{N} \sum_{j=i+1}^{T} \mathbb{E}_q[z_{dn}^{j}] \tag{274}$$

Finally, the indicator random variables ξ_{di} in (269) and latent variables z_{dn} in (270) have variational parameters $\mho_{di}$ and γ_{dn}, respectively. This effectively leads to:

$$\mho_{di}^{k} \propto \exp\{\mathbb{E}_q[\log \varsigma_k] + \sum_{n=1}^{N} \gamma_{dn}^{i} \left(\mathbb{E}_q[\log \varphi_{kv}] + \mathbb{E}_q[\log(\varphi_{k(V+1)})] \right)\}, k \in \{1, 2, ..., K\} \tag{275}$$

$$\gamma_{dn}^{i} \propto \exp\{\mathbb{E}_q[\log \phi_{di}] + \sum_{k=1}^{K+1} \mho_{di}^{k} \left(\mathbb{E}_q[\log \varphi_{kv}] + \mathbb{E}_q[\log(\varphi_{k(V+1)})]\right)\}, i \in \{1, 2, ..., T\} \tag{276}$$

$$\gamma_{dn}^{i} \propto \exp\{\mathbb{E}_q[\log(\phi_{di})]\}, i = T + 1 \tag{277}$$

Deriving variational parameter estimates using expectations of complete conditionals follows the work in [1] on symmetric LDA.

These variational updates and expectations at the corpus and document level DPs characterize a form of HDP-LBLA based coordinate ascent framework when maximizing its lower bound, a scheme that is reminiscent of the EM algorithm where in the E-step, we compute documents local variational parameters while maximizing the ELBO (with the initialized global parameters held fixed). Then in the M-step, we maximize the ELBO with respect to the global parameters given the local context (document's variational parameters).

We express here some useful expectations [167] as shown below:
$\mathbb{E}_q[\xi_{di}^{k}] = \mho_{di}^{k}$
$\mathbb{E}_q[z_{dn}^{k}] = \gamma_{dn}^{k}$
$\mathbb{E}_q\left[\log \varsigma_k^{'}\right] = \Psi(\iota_k) - \Psi(\iota_k + b_k)$
$\mathbb{E}_q\left[\log(1 - \varsigma_k^{'})\right] = \Psi(b_k) - \Psi(\iota_k + b_k)$
$\mathbb{E}_q[\log \varsigma_k] = \mathbb{E}_q\left[\log \varsigma_k^{'}\right] + \sum_{l=1}^{k-1} \mathbb{E}_q\left[\log(1 - \varsigma_l^{'})\right]$
$\mathbb{E}_q[\log \phi_{dt}] = \mathbb{E}_q\left[\log \phi_{dt}^{'}\right] + \sum_{i=1}^{t-1} \mathbb{E}_q\left[\log(1 - \phi_{di}^{'})\right]$
$\mathbb{E}_q\left[\log \phi_{dt}^{'}\right] = \Psi(\vartheta_{di}) - \Psi(\vartheta_{di} + r_{di})$
$\mathbb{E}_q\left[\log(1 - \phi_{di}^{'})\right] = \Psi(r_{di}) - \Psi(\vartheta_{di} + r_{di})$

6.4.3.2 Stochastic optimization for HDP-LBLA with the natural gradient

In our proposed approach, the variational inference accommodates the computation of natural gradients of the ELBO with respect to the variational parameters. In maximum likelihood (ML) framework the natural gradient method is favored as it provides fast convergence than standard gradient (it does not require the whole corpus data to improve global estimates).

The stochastic variational method uses stochastic optimization where it iteratively gets subsamples from the corpus. Doing so, it computes a noisy estimate of the natural gradient of the ELBO which is used to find point estimate of the global variational parameters. With probability measures, the natural gradient scheme characterizes information geometry of the parameter space (the Riemannian space which is the space where local distances are defined by the Kullback Leibler divergence) as it utilizes the Riemannian metric or Fisher information matrix to guide the directions of standard Euclidean gradients [1].

Considering the global topics indexed by the global variational parameters $\tilde{\zeta}$ such that $\tilde{\zeta} = (\tilde{\lambda}_{k1}, ..., \tilde{\lambda}_{kV}, \tilde{\lambda}, \eta)$, we can compute the natural gradient by premultiplying the standard gradient with the inverse of the Riemannian metric $\mathscr{G}$ such that $\hat{\nabla}_{\tilde{\zeta}}\ell(\zeta) \triangleq \mathscr{G}(\tilde{\zeta})^{-1}\nabla_{\tilde{\zeta}}\ell(\tilde{\zeta})$. The metric $\mathscr{G}$ is also called the Fisher information matrix of $q(\varphi|\tilde{\zeta})$, and it is defined as $\mathscr{G} = \mathbb{E}_{\tilde{\zeta}}[(\nabla_{\tilde{\zeta}} \log q(\varphi|\tilde{\zeta}))(\nabla_{\tilde{\zeta}} \log q(\varphi|\tilde{\zeta}))^T]$.

For exponential family variational distribution $q(\varphi|\tilde{\zeta})$, the metric $\mathscr{G}$ becomes the second derivative (the Hessian) of the log normalizer with respect to the variational parameter $\tilde{\zeta}$

such that $\mathscr{G} = \nabla^2_{\tilde{\zeta}} a_g(\zeta)$. The Hessian of the log normalizer with respect to the variational parameter $\tilde{\zeta}$ is also identified as the covariance matrix of the sufficient statistic vectors $t(\varphi)$.

$$\nabla_{\tilde{\zeta}} \ell = \nabla^2_{\tilde{\zeta}} a_g(\tilde{\zeta})(\mathbb{E}_q[\mathscr{C}_g(x, z, \xi, \zeta)] - \tilde{\zeta}) \tag{278}$$

$$\hat{\nabla}_{\tilde{\zeta}} \ell = \mathbb{E}_q[\mathscr{C}_g(x, z, \xi, \zeta)] - \tilde{\zeta} \tag{279}$$

where $\mathscr{C}_g(x, z, \xi, \zeta)$ is the global natural parameter of the complete conditional posterior $p(\varphi_k|z, \xi, x) = BL(\bar{\zeta})$. As we can see in (279), noisy estimates of the natural gradients are simpler to compute than true gradients (278). Importantly, such estimates are often useful as they help algorithms avoiding shallow local optima when optimizing complex objective functions (ELBO). The efficiency of natural gradient techniques allow variational Bayes methods to accommodate large scale applications. We use stochastic optimization with natural gradient method to optimize our variational objective function from SV-HDP-LBLA topic model. Below, we provide our objective function for the stochastic optimization.

$$\begin{aligned}\mathscr{F}(q) = \sum_{j}^{\mathscr{D}} \mathbb{E}_q[\log p(x_j|\xi_j, z_j, \varphi) p(\xi_j|\varsigma) p(z_j|\phi_j^{'})] - \mathscr{H}(q(\xi_j)) - \mathscr{H}(q(z_j)) - \mathscr{H}(q(\phi_j^{'})) \\ + \mathbb{E}\left[\log p(\phi_j^{'}|\varrho)\right] + \mathbb{E}_q\left[\log p(\varphi|\zeta)\right] + \mathbb{E}_q\left[\log p(\varsigma^{'}|\hbar)\right] - \mathscr{H}(q(\varsigma_k^{'})) - \mathscr{H}(q(\varphi))\end{aligned} \tag{280}$$

Our approach is also similar to the work in [182, 1, 65, 167]. With $\mathscr{D}$ defining the total number of documents in the corpus, we divide the corpus wide term by $\mathscr{D}$ such that (265) is rewritten as $\mathscr{F} = \sum_j \mathscr{F}_j = \mathbb{E}_j[\mathscr{D}\mathscr{F}_j]$ from (249) where $\ell = \mathscr{F}_j$. It leads to:

$$\begin{aligned}\ell = \mathbb{E}_q[\log p(x_j|\xi_j, z_j, \varphi)] + \mathbb{E}_q\left[p(\xi_j|\varsigma^{'})\right] + \mathbb{E}_q\left[p(z_j|\phi_j^{'})\right] \\ - \mathscr{H}(q(\xi_j)) - \mathscr{H}(q(z_j)) - \mathscr{H}(q(\phi_j^{'})) + \mathbb{E}_q\left[\log p(\phi_j^{'}|\varrho)\right] \\ + \frac{1}{\mathscr{D}}(\mathbb{E}_q\left[\log p(\varphi|\zeta)\right] + \mathbb{E}_q\left[\log p(\varsigma^{'}|\hbar)\right] - \mathscr{H}(q(\varsigma_k^{'})) - \mathscr{H}(q(\varphi))\end{aligned} \tag{281}$$

The expression above is the variational lower bound for $\mathscr{D}$ copies of document j. The convergence analysis of our HDP-LBLA topic model is similar to the one already discussed in our LBLA model in section 6.3.4.2 due to truncations. With $\mathscr{D}$ copies of document j, we compute the natural gradient estimates of the global variational parameters of the topics:

$$\begin{cases}\hat{\nabla}_{\tilde{\lambda}_{kv}} \ell(\tilde{\lambda}_{kv}) = -\tilde{\lambda}_{kv} + \lambda_{kv} + \mathscr{D} \sum_{i=1}^{T} \mho_{di}^{k} \sum_{n=1}^{N} \gamma_{dn}^{i} x_{dn} \\ \hat{\nabla}_{\tilde{\lambda}} \ell(\tilde{\lambda}) = -\tilde{\lambda} + \lambda + \mathscr{D} \sum_{n=1}^{N} \gamma_{dn} \\ \hat{\nabla}_{\tilde{\eta}} \ell(\tilde{\eta}) = -\tilde{\eta} + \eta + \mathscr{D} \sum_{n=1}^{N} \gamma_{dn} x_{dn}^{(V+1)}\end{cases}$$

The other global variables are the corpus level stick-breaking proportions ς_k where each is associated with a set of Beta distributions.

$$\hat{\nabla}_{\iota_k} \ell(\iota_k) = -\iota_k + 1 + \mathscr{D} \sum_{i=1}^{T} \mho_{di}^{k} \qquad \hat{\nabla}_{b_k} \ell(b_k) = -b_k + \hbar + \mathscr{D} \sum_{i=1}^{T} \sum_{j=k+1}^{K} \mho_{di}^{j}$$

We summarize the local updates: Under the variational method, we compute the latent indicator functions $\mho_{di}$ in (275) and γ_{dn} (276) and (277), and the parameters (ϑ_{di}, r_{di}) of the variational distributions associated to the document-level stick-breaking proportions $q(\phi_{di}^{'}|\vartheta_{di}, r_{di})$ in (282) below.

$$\vartheta_{di} = 1 + \sum_{n=1}^{N} \gamma_{nd}^{i}, i \in \{1, 2, ..., T\} \qquad r_{di} = \varrho + \sum_{n=1}^{N} \sum_{j=i+1}^{T} \gamma_{nd}^{j}, i \in \{1, 2, ..., T\} \tag{282}$$

We update the variational parameters of the global stick-breaking proportions and the global topic parametes $q(\varphi_k|\tilde{\zeta})$ and $q(\varsigma_k^{'}|\iota_k, b_k)$. We recapitulate the global variational parameters updates in (283) and (284).

$$\begin{cases} \hat{\nabla}_{\tilde{\lambda}_{kv}} \ell(\tilde{\lambda}_{kv}) = -\tilde{\lambda}_{kv} + \lambda_{kv} + \mathscr{D} \sum_{i=1}^{T} \mho_{di}^{k} \sum_{n=1}^{N} \gamma_{dn}^{i} x_{dn} \\ \hat{\nabla}_{\tilde{\lambda}} \ell(\tilde{\lambda}) = -\tilde{\lambda} + \lambda + \mathscr{D} \sum_{n=1}^{N} \gamma_{dn} x_{n} \\ \hat{\nabla}_{\tilde{\eta}} \ell(\tilde{\eta}) = -\tilde{\eta} + \eta + \mathscr{D} \sum_{n=1}^{N} \gamma_{dn} x_{dn}^{(V+1)} \\ \hat{\nabla}_{\iota_k} \ell(\iota_k) = -\iota_k + 1 + \mathscr{D} \sum_{i=1}^{T} \mho_{di}^{k} \\ \hat{\nabla}_{b_k} \ell(b_k) = -b_k + \hbar + \mathscr{D} \sum_{i=1}^{T} \sum_{j=k+1}^{K} \mho_{di}^{j} \end{cases} \tag{283}$$

$$\begin{cases} \tilde{\lambda}_{kv} \leftarrow \tilde{\lambda}_{kv} + \rho_t \hat{\nabla}_{\tilde{\lambda}_{kv}} \ell(\tilde{\lambda}_{kv}) \\ \tilde{\lambda} \leftarrow \tilde{\lambda} + \rho_t \hat{\nabla}_{\tilde{\lambda}} \ell(\tilde{\lambda}) \\ \tilde{\eta} \leftarrow \tilde{\eta} + \rho_t \hat{\nabla}_{\tilde{\eta}} \ell(\tilde{\eta}) \\ \iota_k \leftarrow \iota_k + \rho_t \hat{\nabla}_{\iota_k} \ell(\iota_k) \\ b_k \leftarrow b_k + \rho_t \hat{\nabla}_{b_k} \ell(b_k) \end{cases} \tag{284}$$

The proposed online HDP-LBLA which is the SV-HDP-LBLA proceeds as follows, given the corpus level parameters, we first and randomly draw a document j from the corpus and then compute its optimal local context. In other words, we evaluate the document level variational parameters $(\vartheta_j, r_j, \mho_j, \gamma_j)$ using coordinate ascent method in (282), (275), and (276). Then, we take the natural gradient of $\mathscr{D}\mathscr{F}_j$ with respect to the global-level parameters $(\tilde{\lambda}_{k1}, ..., \tilde{\lambda}_{kV}, \tilde{\lambda}, \tilde{\eta}, \iota_k, b_k)$. It is a noisy estimate of the lower bound (ELBO) formulated as an expectation using $\mathscr{D}$ copies of document j. The global variables follow their natural gradients with a step size ρ_t until convergence. When t increases, the step size decreases. In this stochastic method, we could also implement a minibatch framework by using multiple samples (documents) at a time as in the case of SV-LBLA using (255). Furthermore, the minibatch approach ultimately improves the computational speed and the estimates [65].

6.4.4 Time and space complexities

Reaching a model selection, assessing time and memory complexities are some of the most important subjects in topic modeling literature because such terms are related. The work of [12], [80], [18] especially the one in [153] provided a detailed time and memory complexities for standard parametric LDA within the variational inference. Usually, the time complexity of LDA in VB for one iteration is $O(2 \times K \times N0 \times \wp)$, where $\wp$ is the overall time it takes to compute exponential digamma functions. The LDA in general has to handle three types of estimates: two multinomial parameters and latent variables. The document-topic multinomial parameter has a size of $1 \times K$ for each document whereas the topic-word multinomial parameter is a $K \times V$ matrix per document where K is the number of topics while V is the size of the vocabulary. The corpus contains $\mathscr{D}$ documents. Each document has

Table 6.5: Stochastic Variational Inference for Bayesian nonparametric HDP-LBLA

```
Fully asymmetric HDP-LBLA Stochastic Variational Method

◇ Initialize corpus-level HDP-LBLA parameters
◇ Set the learning rate $\rho_t$ ($\rho_t = (\tau + t)^{-\kappa}, t = 1, t \leftarrow t + 1$)
◇ Select a document $d$ uniformly from the corpus data $\mathscr{D}$
        ◇ E-STEP
◇ Update document local variational parameters: $\vartheta_{di}, r_{di}, \mho^k_{di}, \gamma^i_{dn}$
using (274) or (282), (275), and (276)

        ◇ M-STEP
◇ Compute noisy estimates of the natural gradients:
$\hat{\nabla}_{\tilde{\lambda}_{kv}} \ell(\tilde{\lambda}_{kv}), \hat{\nabla}_{\tilde{\lambda}} \ell(\tilde{\lambda}), \hat{\nabla}_{\tilde{\eta}} \ell(\tilde{\eta}), \hat{\nabla}_{\iota_k} \ell(\iota_k), \hat{\nabla}_{b_k} \ell(b_k)$
with $\mathscr{D}$ copies of document $d$ using (283)
◇ Compute global parameters with natural gradients with (284)
($\tilde{\lambda}_{kv}$, $\tilde{\lambda}$, $\tilde{\eta}$, $\iota_k$, and $b_k$)
◇ Until convergence
◇ Output: $\tilde{\lambda}_{kv}$, $\tilde{\lambda}$, $\tilde{\eta}$, $\iota_k$, and $b_k$
```

$N0$ (nonzero) elements in document-word sparse matrix. The corpus has therefore a size of $\mathscr{D} \times V$. The total space complexity of VB-based LDA is around $O(\mathscr{D} \times N0 + 2 \times K \times (\mathscr{D} + V))$. The HDP-LDA in VB inference, has to perform five estimates which include the corpus breaking proportions, the document-level topic indices, the global topic, the document-level breaking proportions, and the latent variables. This automatically increases the memory requirement of the HDP-LDA compared to standard LDA. The memory complexity of the HDP-LDA is around $O(2 \times \mathscr{D} \times K + 2 \times V \times T + N0 \times T + \mathscr{D} \times N0)$.

As expected, the HDP-LDA is slow compared to standard LDA with a time complexity usually higher than that of LDA. Our truncated HDP-LBLA has a time and memory complexities almost similar to that of HPD-LDA. However, it is relatively faster than HDP-LDA. Its AA structure leads to very flexible stick-breaking priors (probability measures) that provides very heterogeneous topics which allow the proposed approach to select quickly relevant topics while discarding the irrelevant ones. The truncated HDP-LDA with the standard symmetric Drichlet prior has no ability to provide such flexibilities in terms of constructing fast and relevant topics. Naturally, we expect the HDP-LBLA to be slower than the standard LDA because of the high complexity in BNP. However, its estimates (predictive distributions) are accurate compared to the standard (parametric) LDA.

Convergence is not an issue in variational Bayes as the models are deterministic methods. It is important to mention that while implementing a stochastic (online) method here, the time and memory complexities characterized in this section only concern the batch techniques. It is implicitly clear that in online scheme, the number of documents accessed at a time is much reduced ($S << \mathscr{D}$); therefore, with an improved time and memory complexities, our online methods are much faster to reach convergence or providing estimates. This suggests that for large scale processing, we expect the batch HDP-LBLA to be slower than online versions despite its flexibility in providing accurate number of topics. Particularly, while the online HDP-LBLA operates in infinite dimensional space within a countably infinite number of topics, it reaches a model selection with fewer number of topics. It means that in online HDP-LBLA, time and memory complexities components such as K, T with ($K >> T$), and S are much smaller; therefore, its has much better time and memory complexities than batch HDP-LBLA.

6.5 Experimental Results

Our experimental section is based on three datasets for text document analysis. It includes these large scale datasets to take advantage of our SV-HDP-LBLA also called online HDP-LBLA (OHLB).

6.5.1 Implementation

In implementation, we have initialized the global variational parameters. This is a stochastic mean-field variational inference at document level. Only the global parameters are recorded while the local context is discarded including the latent variables. The goal is to optimize (maximize) our objective function, the ELBO, with respect to the variational parameters. The framework requires an initial setting concerning the model's hyperparameters. We usually set them randomly. However, for the BL hyperparameters, we could also provide initializations as follows: at the document level we choose $\alpha_{dt} = \frac{1}{t}$ where $t \in \{1, 2, ..., T\}$ to characterize asymmetric BL prior. We also set α_d such that $\alpha_d < \sum_{t=1}^{T} \alpha_{dt}$ or $\alpha_d > \sum_{t=1}^{T} \alpha_{dt}$. Then, we choose the value of β_d within the same scale as α_d ($\beta_d > \alpha_{d(T+1)}$). At the corpus level for BL, we repeat the same process by setting values for λ_{kv} with $v \in \{1, 2, ..., V\}$ and λ and η where the truncation is set at K ($k \in \{1, 2, ..., K\}$). In our truncated HDP-LBLA's implementation $T << K$.

While our implementation is also described in section 6.4.3.2, we could summarize it here as follows: given the corpus level parameters, we randomly draw a document j from the corpus and then compute its local parameters which means that at the E-step (expectation), we evaluate the document level variational parameters $(\vartheta_d, r_d, \mho_d, \gamma_d)$ using coordinate ascent method in (282), (275), and (276). Then, we compute estimate of the natural gradient using the ELBO corresponding to $\mathscr{D}$ copies of document d ($\mathscr{D}\mathscr{F}_d$) with respect to the global-level parameters $(\tilde{\lambda}_{k1}, ..., \tilde{\lambda}_{kV}, \tilde{\lambda}, \tilde{\eta}, \iota_k, b_k)$ in M-step (maximization). We also provided an option of a minibatch framework using multiple samples (documents) at a time as in the case of SV-LBLA using (255) to improve both estimates and computational speed. At convergence, the global parameters are estimated. They are point estimates. In case minibatch sizes become too small and affect performamce, we can provide much larger minibatch sizes. In these experiments we use minibatch sizes such that $S = \{10, 50, 200, 500, 1000\}$.

We set the step size ρ_t (learning rate) at iteration t such that $\rho_t = (t + \tau)^{-\kappa}$. The forgetting rate $\kappa \in (.5, 1]$ controls how quickly old information is forgotten, during successive iterations. The algorithm is guaranteed to converge at least to a local optimum of the ELBO. The method is summarized in Table 6.5.

6.5.2 Datasets

To show the flexibility and performance of our proposed approach, we used three challenging datasets in text document processing. These collections are the NIPS dataset, KOS text documents, and ENRON text data as shown in Table 6.6. The NIPS dataset is a collection from scientific papers from the proceedings of NIPS database. It has roughly around 2484 papers. The corpus contains $\mathscr{D} = 1740$ documents for a total vocabulary size of $V = 12419$. It also carries a total of $N = 2166029$ words and $M = 836644$ unique word-document pairs. The KOS collection is from the report blog website (online). It has a total of $\mathscr{D} = 3430$ documents, a vocabulary size of $V = 6909$, and a total of $N = 467714$ words and $M = 360664$ unique word-document pairs. The ENRON dataset has total corpus in documents of $\mathscr{D} = 39861$; with a vocabulary size of $V = 28102$, it provides a total of $N = 6400000$ words.

Table 6.6: Text document datasets

	$\mathscr{D}$train	$\mathscr{D}$test	N	V	$\mathscr{D}$
NIPS	1256	419	2166029	12419	1675
KOS	2573	857	467714	6909	3430
ENRON	29896	9965	6400000	28102	39861

6.5.3 Methodology and evaluation method

The text documents are represented using the Bow (bag of words) approach. We divide each collection into two parts: 80% of the corpus documents for training and 20% for testing (held-out). We used the training set to estimate the model parameters especially φ_k. For evaluation, we use the likelihood per word method as in [167, 1, 16, 4]. For testing or validation, each document d containing $\boldsymbol{x}_d$ words is then divided into two parts: $\boldsymbol{x}_{d1}$ and $\boldsymbol{x}_{d2}$. The first part $\boldsymbol{x}_{d1}$ takes 80% of the total words in $\boldsymbol{x}_d$ while the second part $\boldsymbol{x}_{d2}$ holds the 20% in words in document d. In topic modeling, the parameter of the multinomial distribution of the topic assignment z_{dn} represents the document topic proportion. The topic proportion linked to a document d is K-dimensional while the corpus parameter φ_k is V-dimensional.

In our HDP-LBLA model, $z_{dn} \sim Mult(\phi_{dt})$, therefore $\phi \sim GEM(\varrho)$ and $\phi = \{\phi_{dt}\}_{t=1}^{\infty}$ which represent the weights (stick lengths) obtained from the stick-breaking (proportions) method at the document level. However, ϕ_d is infinite-dimensional, and the global topic $\varphi = \{\varphi_k\}_{k=1}^{\infty}$ is also infinite dimensional. The atoms at the document level Ω_{dt} and the corpus level φ_{kv} coincide at $\Omega_{dt} = \varphi_{\xi_{dt}}$ (with $\xi_d = \{\xi_{dt}\}_{t=1}^{\infty}$) meaning at locations pointed by ξ_{dt}. Therefore, when the global topic is truncated at K, only a maximum of K global topics can be shared at the document-level DP. Within the variational framework, these parameters φ_k and ϕ_d become expectations with respect to their variational distributions. Precisely, as in [167], $\mathbb{E}_q[\varphi]$ is the variational expectation using the training data $\mathscr{D}_{train}$; and $\mathbb{E}_q[\phi_d]$

is the variational expectation using $\boldsymbol{x}_{d2}$ which is a predictive model as we keep the global $\mathbb{E}_q[\varphi]$ fixed during testing. Therefore, the predictive distribution of $\boldsymbol{x}_{d2}$ is approximated using [167, 1]. We define the predictive likelihood $p(\boldsymbol{x}_{d2}|\mathscr{D}_{train}, \boldsymbol{x}_{d1})$ for $\boldsymbol{x}_{d2}$ documents as $p(\boldsymbol{x}_{d2}|\mathscr{D}_{train}, \boldsymbol{x}_{d1}) = \prod_{x \in \boldsymbol{x}_{d2}} \int\int \left(\sum_{k=1}^{K}(\phi_{dk})\varphi_{kx}\right) p(\phi|\boldsymbol{x}_{d1}, \varphi)p(\varphi|\mathscr{D}_{train})d(\phi)d\varphi$ which is approximated as follows:

$$p(\boldsymbol{x}_{d2}|\mathscr{D}_{train}, \boldsymbol{x}_{d1}) \approx \prod_{x \in \boldsymbol{x}_{d2}} \int\int \left(\sum_{k=1}^{K} \phi_{dk}\varphi_{kx}\right) q(\phi)q(\varphi)d(\phi)d\varphi \tag{285}$$

$$\approx \prod_{x \in \boldsymbol{x}_{d2}} \sum_{k=1}^{K} \mathbb{E}_q[\phi_{dk}]\mathbb{E}_q[\varphi_{kx}] \tag{286}$$

where $\mathcal{L}$ is the per-word predictive likelihood in $\boldsymbol{x}_{d2}$ documents.

$$\mathcal{L} \approx \frac{\sum_{d \in \mathscr{D}_{test}} \log p(\boldsymbol{x}_{d2}|\boldsymbol{x}_{d1}, \mathscr{D}train)}{\sum_{d \in \mathscr{D}test} |\boldsymbol{x}_{d2}|} \tag{287}$$

such that $|\boldsymbol{x}_{d2}|$ is the cardinality of $\boldsymbol{x}_{d2}$ or the total number of words in $\boldsymbol{x}_{d2}$.
When computing the predictive distribution for the document parameter, in G_d every topic assignment is always matched to a draw from the base measure. When the base measure is marginalized out, new topic assignment will be matched to a draw using the global distribution [9]. In empirical Bayes framework for hyperparameter estimation, integrating out the base measure at the document-level DP makes the $\varsigma \sim GEM(\hbar)$ act as infinite dimensional hyperparameters for the prior used.

6.5.4 Predictive models for HDP-LBLA

We aim to provide the prediction rule for estmating the probability of previously unseen documents. Since $G_d \sim DP(\varrho, G_0)$ and $\varsigma \sim GEM(\hbar)$ and $\phi \sim GEM(\varrho)$, then $\phi_d \sim DP(\varrho, \varsigma)$. We compute these identities from the Sethuraman's stick-breaking method to facilitate estimation of predictives model in our HDP-LBLA topic model.
$\mathbb{E}_q[\log \phi_{dt}] = \mathbb{E}_q\left[\log \phi'_{dt}\right] + \sum_{i=1}^{t-1} \mathbb{E}_q\left[\log(1 - \phi'_{di})\right]$
$\mathbb{E}_q\left[\log \phi'_{dt}\right] = \Psi(\iota_k) - \Psi(\vartheta_{di} + r_{di})$
$\mathbb{E}_q\left[\log(1 - \phi'_{di})\right] = \Psi(r_{di}) - \Psi(\vartheta_{di} + r_{di})$

The Beta random weights are defined as:

$$\mathbb{E}_q\left[\phi_{dk}|\vartheta_{dk}, r_{dk}\right] = \frac{\vartheta_{dk}}{\vartheta_{dk} + r_{dk}} \prod_{l=1}^{k-1} \frac{r_{dl}}{\vartheta_{dl} + r_{dl}} \tag{288}$$

where the Beta parameters ϑ_{di} and r_{di} are computed using (282). We can deduct the expectation of the corpus stick-breaking proportions:

$$\mathbb{E}_q\left[\varsigma_k|\iota_k, b_k\right] = \frac{\iota_k}{\iota_k + b_k} \prod_{l=1}^{k-1} \frac{b_l}{\iota_l + b_l} \tag{289}$$

with the corpus Beta parameters ι_k and b_k computed using (273). At the top level DP, the global probability measure G_0 varies around the expected value H which is a BL (non atomic measure) distribution with a concentration parameter $\hbar$. Similarly, the local measure G_d varies around the expected value G_0 with a degree of variability controlled by the concentration parameter ϱ. Our GEM variables (random weights) are $\phi = \{\phi_{dt}\}_{t=1}^{\infty}$ and $\varsigma = \{\varsigma_k\}_{k=1}^{\infty}$. At infinite limit of finite mixtures, we have to set the document BL parameters such that:

$$\begin{cases} \alpha_k & = \varrho \mathbb{E}_q[\varsigma] = \varrho \times \frac{\iota_k}{\iota_k + b_k} \prod_{l=1}^{k-1} \frac{b_l}{\iota_l + b_l} \\ \alpha & > \sum_{k=1}^{K} \alpha_k \quad or \quad \alpha < \sum_{k=1}^{K} \alpha_k \\ \beta & > \alpha_{K+1} \quad or \quad \beta < \alpha_{K+1} \end{cases} \tag{290}$$

where $k \in \{1, 2, ..., K\}$. We get a Dirichlet as special case over the finite partitions with parameters α_k such that:

$$\alpha_k = \varrho \times \frac{\iota_k}{\iota_k + b_k} \prod_{l=1}^{k-1} \frac{b_l}{\iota_l + b_l}$$

where $k \in \{1, 2, ..., K\}$ and its weights are Beta random variables from the GEM. In that case, the truncated GEM will approximate the HDP-LDA (infinite limit of finite topic mixtures). The variational posterior is given as:

$$\tilde{\alpha}_{dk} = \varrho \mathbb{E}_q[\varsigma] + \sum_{n=1}^{N} \gamma_{dn}^{k} \qquad \tilde{\alpha} = \alpha + \sum_{n=1}^{N} \sum_{k=1}^{\mathscr{D}} \gamma_{dn} \qquad \tilde{\beta} = \beta + \sum_{n=1}^{N} \sum_{d=1}^{\mathscr{D}} \gamma_{dn}^{K+1} \tag{291}$$

We see the relationship between the Dirichlet and the BL prior. The BL has just two more parameters than the Dirichlet; yet it has much flexibility over the Dirichlet with a more general covariance structure (235). Therefore, to use the approximation in (286), we implement (290) and (291). The truncated GEM in our case will approximate the HDP-LBLA. As in section 6.3.4.3, we estimate $\mathbb{E}_q[\varphi_{kx}]$ from the training set and keep it fixed where we compute $\mathbb{E}_q[\phi_{dk}]$.

6.5.5 Experiments and Results

In this chapter, because the LBLA generalizes the LDA, we decided to compare directly our BL-based HDP topic model to its parametric counterpart which is the LBLA. The goal is to show the performance of these two models under different datasets. We will have to predefine the number of topics for the finite dimensional parametric LBLA whereas for the HDP, we will let the model choose its components based on the data. In other words the HDP-LBLA will be initially set to the maximum of topics. We will also show how the stochastic variables such as the batch size S and the forgetting rate κ influence the performance. We will also show how different number of topics affect the parametric LBLA. Topic models in general have to estimate K topics from a vocabulary of size V per iteration using the training data. Batch methods require all the training dataset while stochastic online variational methods only utilize subsets (minibatches). Therefore, we will assess the per-iteration performance of these models (batch an online schemes) in terms of their predictive likelihood per document.

Despite the flexibility of the perplexity method as an evaluation scheme it has been reported that the predictive likelihood framework is an alternative as it avoid comparing bounds. As evaluating the predictive distributions avoids comparing bounds, we chose the per-word log predictive probability as an evaluation method. [1], [167]. Below is our initial setting concerning the size of the topics K and minibatch S: $K \in \{10, 20, 40, 50, 60, 100, 120, 180, 200\}$. The minibatch size is selected from $S \in$

$\{50, 100, 200, 500, 1000\}$. We maintained the second level truncation to $T = 10$ following the work in [1, 167] where $K >> T$.

In our figures, the online LBLA (the online SV-LBLA) is the OLB to keep it short. The HLB is the batch-HDP-LBLA (the nonparametric hierarchical topic model), and finally the OHLB is the online version of HLB (the online SV-HDP-LBLA). Throughout the experiments, we use these acronyms interchangeably.

6.5.5.1 NIPS dataset

The per-iteration step observed in online HDP-LBLA and online parametric LBLA (OLB) is much improved compared to that of the batch HDP-LBLA as we notice high value in their log predictive probability for a held out document. The per-iteration step of the batch HDP (HLB) is much slower as it gets penalized with a low value in its per-word log predictive probability. The batch method when using variational inference only follows the coordinate ascent framework which requires the use of all the training data in order to estimate the global parameters such as topics. The time and memory complexities we provided in subsection 6.4.4 show the batch is slower. Furthermore, batch methods are easy to get stuck at local optima, and this could potentially affect the per-word log predictive likelihood. Stochastic online variational schemes can easily escape shallow local optima as they use noisy estimates of the natural gradients [1].

These reasons explain the high performance of our proposed online methods compared to the batch in Figs. 6.1 and 6.2 as they maintain a constantly high value in the log predictive likelihood over the batch at each iteration step. Though, these figures also show that online HDP-LBLA outperforms the online parametric LBLA (OLB). It implicitly shows that online HDP-LBLA has a much bigger hypothesis space than OLB which provides its flexibility when complexity rises in the model. Our parametric models have a very reduced hypothesis space and could not perform efficiently when for instance the number of topics rises. For an extremely high number of topics, they tend to overfit. The BL-based HDPs are much robust: though, while the OHLB is faster, the HLB (batch) is the slowest from these three models due its inefficient time and memory complexities. For a moderate size of topics (Figs. 6.3a and 6.3b), we see that the parametric online model's performance (the per-word log predictive probability value) increases with the size of topics. However, the nonparametric OHLB outperforms both the online stochastic parametric and the batch methods. The online schemes (parametric and nonparametric) that we propose here seem to favor larger batch size along with a slower forgetting rate. This observation is similar to work in [182], [1].

In this dataset, we maintain a $\kappa = 0.7$ as it allows good convergence. The forgetting rate monitors how fast old information are forgotten. This stochastic framework also includes the delay variable $\tau = 1$ (in our experiment). The online HDP obtained the optimal number of topics for the NIPS dataset using $S = 100$ at $K = 70$ from the initial maximum of $K = 200$ topics. We can therefore observe that this dataset favors a small number of topics. Importantly, we could mention the heterogeneity in topics at $K = 50$ in Fig. 6.3c. The GEM structure of the HDP allows such enhanced variability. This promotes a fast detection of relevant topics and an alternative to model selection. It is possible for the parametric online model under a good initial guess on the number of topics to yield also a model selection. Despite the flexibility of the BL-based HDP in determining automatically the number of topics, the performance of our online parametric LBLA is almost close to that of the online HDP-LBLA as shown in Fig. 6.2. Though, without an initial good guess

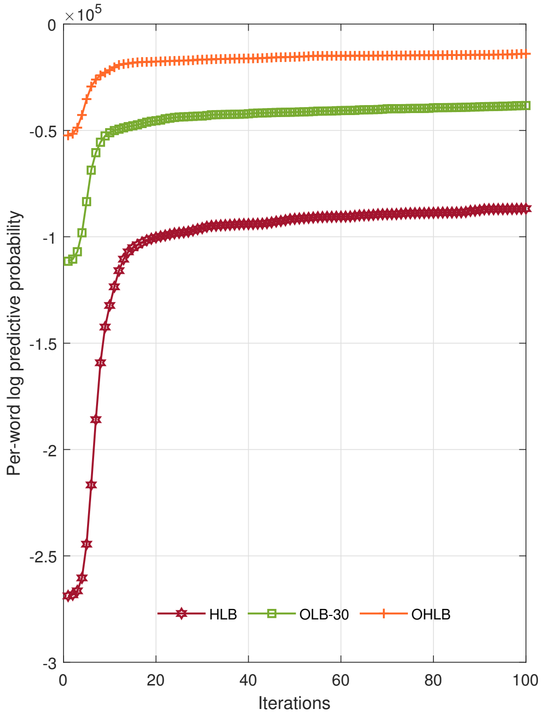


Figure 6.1: Per-word log predictive probability from NIPS dataset showing the performance of the batch HDP-LBLA (HLB), online-LBLA set with 30 topics, and online HDP-LBLA (OHLB).

on the number of topics, the online HDP-LBLA has always the edge over parametric topic models. This essentially represents the central problem with parametric topic models for their inability to assess efficiently the optimal number of topics. As a result, they heavily rely on predefining the exact number of topics which is actually unknown.

6.5.5.2 ENRON dataset

Experiments with this dataset show immediate dominance of online HDP-LBLA over its parametric competitor, the online parametric LBLA as illustrated is Figs. 6.4 and 6.5, especially in Fig. 6.5. The per-iteration step (performance) in the proposed online HDP-LBLA (OHLB) shows that the model estimates the per-word log predictive likelihood faster than its online parametric counterparts. Compared to NIPS data, we notice a much larger batch size for the stochastic framework which also favors a slower forgetting rate κ.

We set the batch size to $S = 500$ while $\kappa = 0.7$. The efficiency of online the BL-based HDP could be explained by the stick-breaking scheme that provides heterogenous topics where we can characterize relevant topics by assessing the probability masses associated to topics [18]. This makes the model robust when the number of topics increases as shown in Figs. 6.6a, 6.6b, and 6.6c. Importantly, Fig. 6.6c shows very heterogenous topic features. Characterizing relevant topics through the GEMs within a stochastic online variational

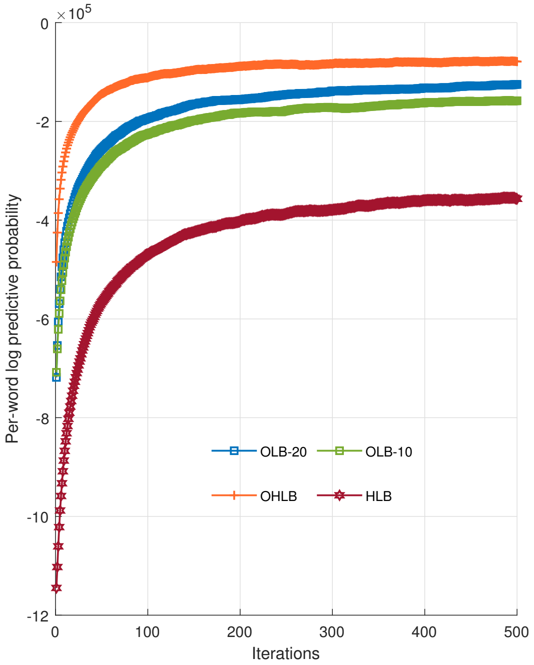
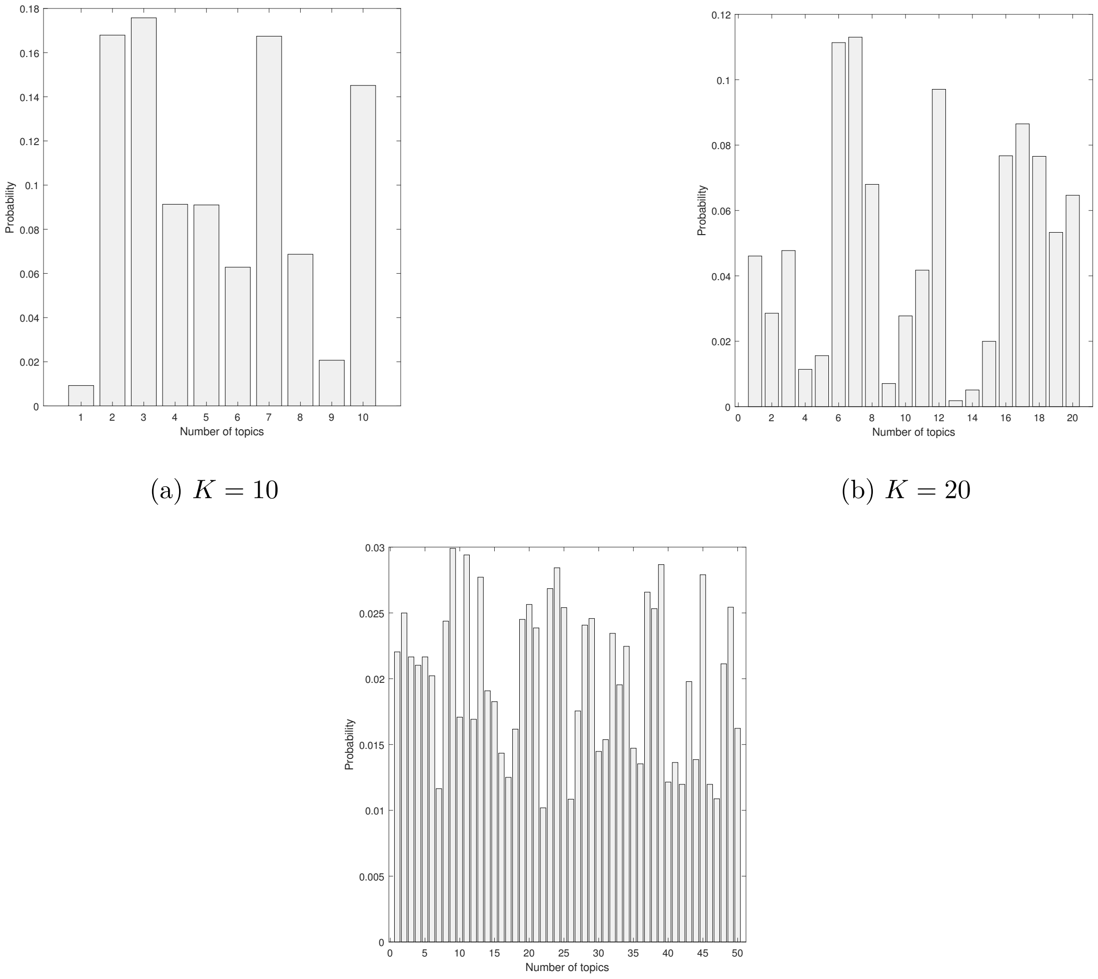


Figure 6.2: Per-word log predictive probability from HDP-LBLA (HLB), online parametric LBLA with 10 and 20 topics (OLB-10 and OLB-20), and online BL-based HDP (OHLB). Performances obtained using NIPS data.

scheme leads to improvement in the predictive estimates. Furthermore, the (online) HDP-LBLA) learns quickly and updates its local and global parameters faster than the batch technique in Fig. 6.5. This flexibility allows it to constantly monitor and maintain an increased value in its per-word log predictive probability (for a held out document) when observing its performance per iteration step. Its competitors do not have such ability. The online parametric model OLB only relies on its stochastic nature to perform quickly as it is not equipped with GEM random variables whose efficient truncations help in the quest for an alternative to model selection. Its performance increases with a good guess on the number of topics; though, OLB is outperformed by online HDP-LBLA. We can see that the performance of the proposed online BL-based HDP is somehow the asymptote for parametric topic models in terms of the high value in log predictive probability (per word) in a document per-iteration (Figs. 6.4 and 6.5).

The batch needs speed to maximize its performance (with predictive likelihood) as fast as its online counterparts. However, the possibility of getting stuck at some local optima remains one of the main reasons the batch method does not provide a good estimate when analyzing its per-word log predictive probability distribution despite being slow. This is not a problem for our stochastic online methods (Figs. 6.4 and 6.5) that follow the noisy estimates of the natural gradients (to escape shallow local optima when the objective function is complex) [1, 157]. They ultimately follow the gradients with a decreasing learning rate. The online parametric LBLA seems to handle efficiently small number of topics as they increase its performance in predictive likelihood (Figs. 6.6a and 6.6b). In addition, despite its noticeable and increased performance within an increased number of topics(with



(a) $K = 10$

(b) $K = 20$

(c) $K = 50$

Figure 6.3: Variability in the topic structure with NIPS dataset (where K is the number of topics).

$K = 30$ and $K = 40$), online parametric topic models (OLB) are still limited compared to our proposed online HDP-LBLA. From the 200 topics, the online HDP-LBLA provides an optimal $K = 100$ topics for $S = 500$ and $\kappa = 0.7$.

6.5.5.3 KOS dataset

We set the batch size to $S = 200$. The KOS dataset maintains a good performance with online HDP-LBLA topic model. We also initialized the parametric topic model with a high number of topics to compete with online BL-based HDP. It results in predictive models that were not desirable because the model is overfitting. The reduced hypothesis space of the parametric model surely does not allow it to accommodate large number of topics. In other words, the online parametric LBLA or OLB becomes unstable for a highly increased number of components. While the performance of the online parametric LBLA was almost similar to

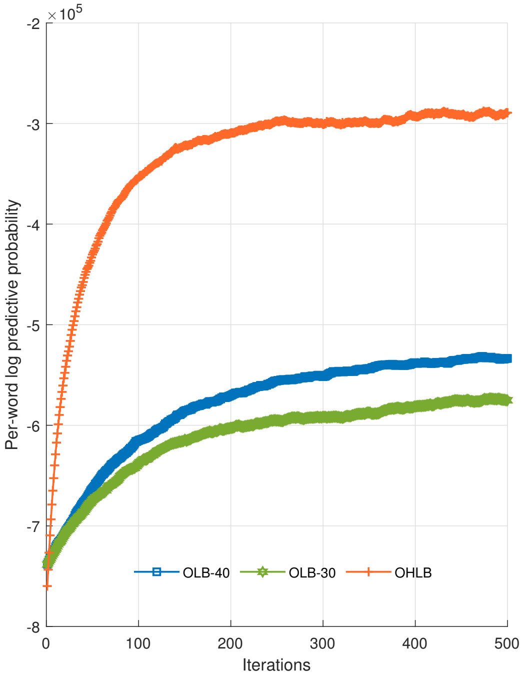


Figure 6.4: Performance of online parametric-LBLA (OLB) using 30 and 40 topics (OLB-30 and OLB-40), and online HDP-LBLA (OHLB) in terms of per-word log predictive probability using ENRON data.

that of the online nonparametric for the NIPS dataset as shown in Fig. 6.2, the online HDP-LBLA in this experiment with KOS dataset clearly outperforms any parametric topic model. As previously mentioned, several attempts to overload the model with a highly increased number of topics could negatively affect the per-word log likelihood estimates. This suggests the parametric models for this dataset require not so high number of topics (Figs. 6.9a, 6.9b). The performance per iteration of online HDP-LBLA in Figs. 6.7 and 6.8 could also be explained by the fact that it favors a larger batch size than in NIPS dataset. As a result, the model could provide much better estimate with its stochastic variational inference. Again the lack of stochasticity in the batch HDP-LBLA makes the model behaving really slow. Within even one iteration, the online models including parametric ones outperform the batch because they do not require the use of all the training data to provide updates. Therefore, the online versions offer much improved predictive estimates per iteration resulting in the high value in the per-word log predictive likelihood shown in Figs. 6.7 and 6.8. They do so faster and efficiently than the batch method. With its flexibility to operate effectively with heterogeneous topics Fig. 6.9c, the online HDP-LBLA reach an optimal number of topics with $K = 80$ and $\kappa = 0.7$.

6.5.5.4 Discussion

Before the general conclusion in section 6.6, from all these three experiments, we can make some reasoning that the performance of the batch HDP-LBLA does not necessarily mean

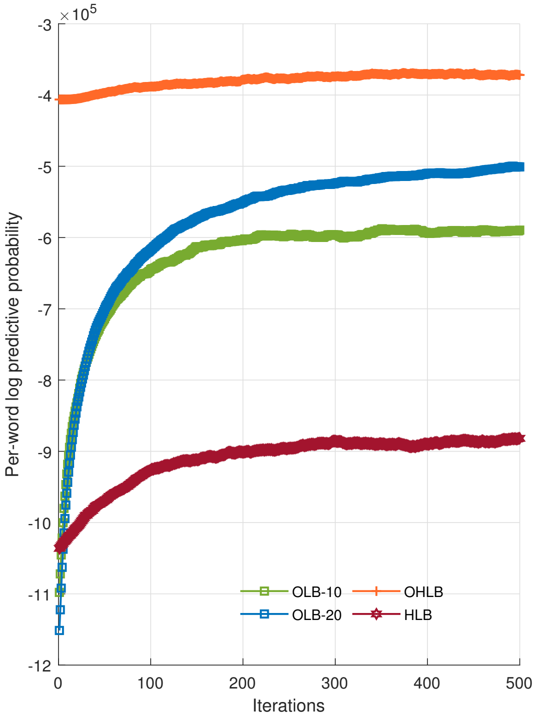
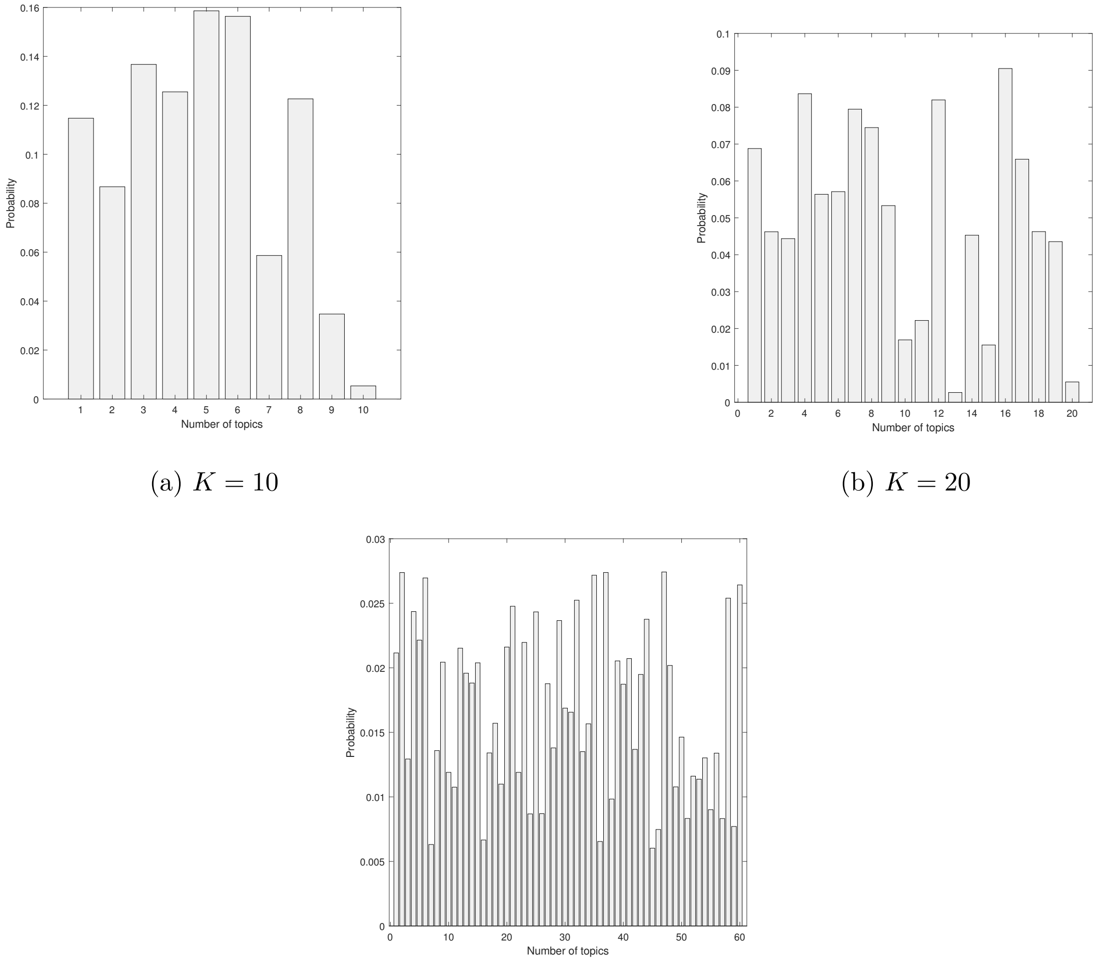


Figure 6.5: ENRON data showing the per-word log predictive probability's performance of batch HDP-LBLA (HLB), parametric online-LBLA using 10 and 20 topics (OLB-10 and OLB-20), and online BL-based HDP (OHLB).

that batch techniques are not useful at all. For large scale applications where $\mathscr{D} \to \infty$, the batch is severely penalized both in time and memory complexities as illustrated in subsection 6.4.4. It makes the batch schemes, despite their capabilities, not efficient compared to sophisticated online methods (based on minibatches of size S such that $S << \mathscr{D}$) which can use the noisy estimates of the natural gradient of the objective function to escape local optima. This flexibility allows online methods to be more efficient in providing estimates faster than batch methods that require the whole training dataset at their disposal at every single iteration.

When speed matters, online methods should be favored than batch framework because they provide efficient estimates more rapidly than any batch system. However, when very accurate results are needed and speed is not required, the batch methods could be used because as they utilize the entire dataset, they can fully uncover the intrinsic structure of the data. Though, their performance per iteration will be always lower: as they need the whole data to provide estimates they are slow while online methods are capable to compute through minibatches several estimates within one iteration. This allows online approaches to reach maximum likelihood estimate faster. By using minibatches, online methods are much efficient and faster. They can efficiently summarize the data characteristics. In this chapter, we show that our online HDP-LBLA, in large scale applications is the right candidate for learning topic models as it has ability to learn the underlined number of components that describes the data efficiently. Online parametric topic models do not have such flexibility, in general. The results in the experiments show that online HDP-LBLA outperforms its



(a) $K = 10$

(b) $K = 20$

(c) $K = 60$

Figure 6.6: Heterogeneous topics from ENRON data as the size of topic increases

batch version and the parametric models.

6.6 Conclusion

we address, in this chapter, the problem of model selection and the sharing ability of clusters in parametric topic models (mixed-membership models). We proposed the HDP with the BL distribution (diffuse base probability measure) as the conjugate prior to the data distribution. Through our experiments with three challenging text datasets, we show the performance of the proposed nonparametric topic model against its parametric counterpart. We assess the performance by averaging the predictive log likelihood within a held out document leading to the per-word predictive log probability value as an evaluation method. We consider it as an alternative to perplexity for its widely use in text document modeling. We showed that the proposed online methods clearly outperform the batch HDP-LBLA technique. The variational coordinate ascent framework, which requires the

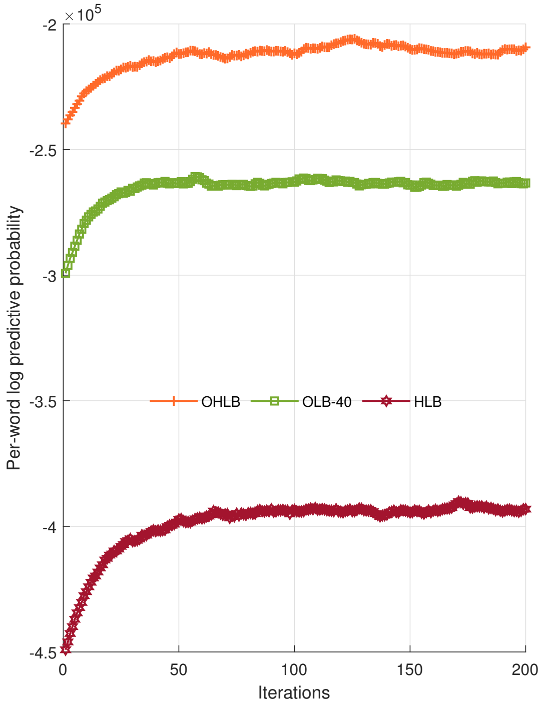


Figure 6.7: From KOS data, the per-word log predictive probability shows the performance of the batch HDP-LBLA (HLB), online parametric LBLA (OLB) using 40 topics, and online HDP-LBLA (OHLB).

batch HDP-LBLA method to use all the available training dataset before updating its variational parameters is not efficient. It makes the batch HDP-LBLA technique very slow and inefficient when we also analyze its time and memory complexities. Because the batch approach uses all the available data, it is more likely that it also gets stuck at some shallow local optima due to the lack of stochasticity in its structure. When it gets stuck, it affects its likelihood value. We observed such phenomenon throughout our experiments. On the other hand, online methods have shown high performance with the per-word log predictive probability value per-iteration. At every iteration, their likelihood is constantly maintained larger than that of the batch HDP-LBLA. These experiments ultimately show the flexibility of the natural gradient method over classical gradients as it characterizes the information geometry of the parameter space and therefore allows a much better estimate.

Online methods demonstrated that they could offer estimate faster with their minibatch scheme. As a result, they have a much improved time and memory complexities. In NIPS dataset, the online HDP-LBLA and parametric LBLA almost perform similarly. It demonstrates that sometimes some good initializations (number of topics) can improve performance even though this is not efficient because in general the number of topics is unknown for a parametric topic model. The parametric online LBLA shows that its performance constantly depends on the number of topics. Its predictive likelihood seems to increase with an increase in the number of topics until it overfits. While the performance of online parametric LBLA was almost close to that of online HDP-LBLA model in NIPS dataset, it is outclassed in the remaining datasets by online HDP-LBLA.

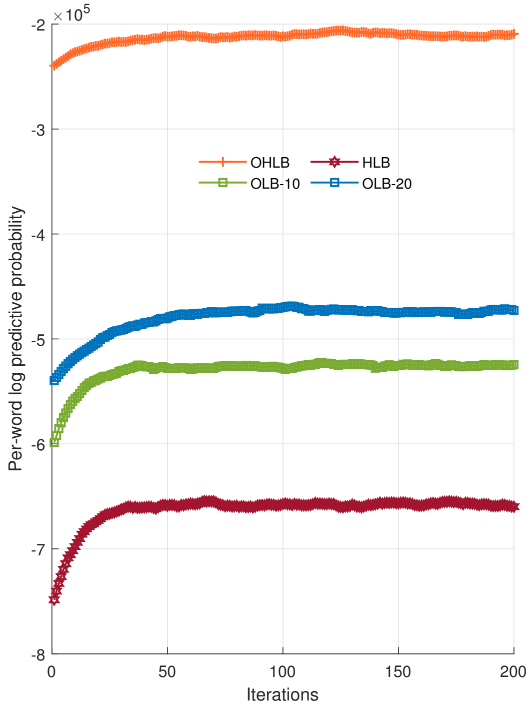
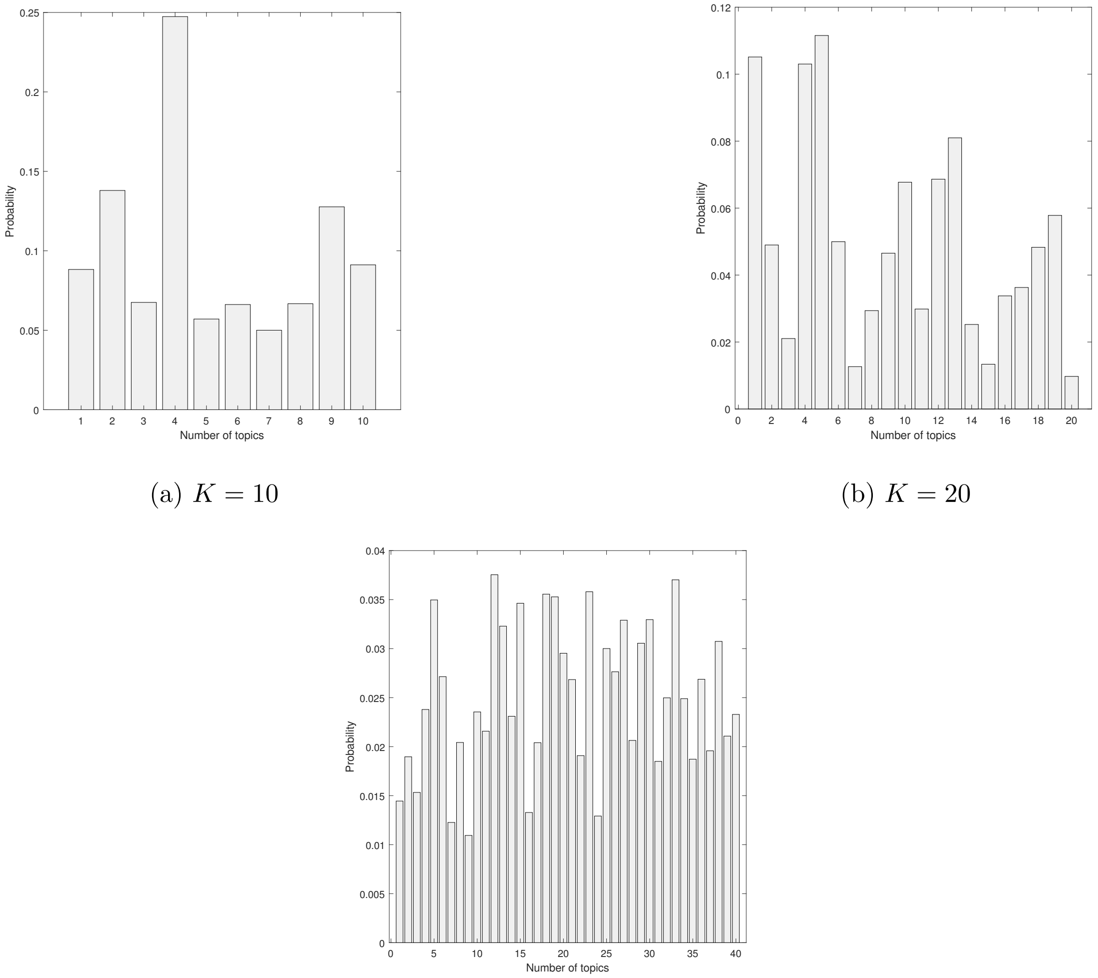


Figure 6.8: Per-word log predictive probability of the batch HDP-LBLA (HLB), online-LBLA with 10 and 20 topics (OLB-10 and OLB-20), and online HDP-LBLA (OHLB) from KOS data

Despite its bigger hypothesis space, the high performance of the HDP-LBLA is due to the heterogeneity of its latent clusters which allows the model to assess rapidly relevant topics. The online HDP-LBLA has a robust GEM structure and BL (asymmetric prior) that offer an alternative to model selection. We also have witnessed that our online models (nonparametric and parametric) tend to favor larger batch size with slow forgetting rate for better estimates. Furthermore, the HDP-LBLA can handle the maximum of topics. Marginalizing over the parameters shows the clustering property (partition) in the dataset. This is reminiscent of the Polya urn process. The efficiency in the predictive log likelihood could be explained by the robustness in the compression algorithm in HDP-LBLA which emphasizes on dependency between documents as they share topics. The ability of our model to characterize topic correlation also explains this flexibility. Even though our online methods outperform the batch technique, the online HDP-LBLA clearly outperforms both the online parametric LBLA and batch HDP-LBLA. We consider it as the most versatile online BL-based HDP topic model. A future work could be devoted to providing another alternative to the HDP-LDA topic model with the generalized Dirichlet distribution as the conjugate prior to the document multinomials.



(a) $K = 10$

(b) $K = 20$

(c) $K = 40$

Figure 6.9: An example of topic structure observed from KOS data at different resolutions

Appendix

Coordinate ascent and update equations

We formulate the coordinate ascent method from the variational inference where we obtain the updates equations. These updates at the corpus and document levels will be useful for the implementation of the parametric and nonparametric topic nodels especially during the training and testing phases.

$$\mathscr{L}'(\gamma_{nk}) = (\mathbb{E}_q[\log \varphi_{kv}] + \mathbb{E}_q[\log(\varphi_{k(V+1)})]) + (\mathbb{E}_q[\log \theta_{dk}] - \log \gamma_{nk} - 1) + \lambda$$

Setting the derivative equal to zero, and we obtain:

$$\gamma_{nk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \varphi_{kv}] + \mathbb{E}_q[\log(\varphi_{k(V+1)})]\} \tag{292}$$

$$\gamma_{nk} \propto \frac{\exp\{\Psi(\tilde{\alpha}_k) + \Psi(\tilde{\alpha})\}}{\exp\{\Psi(\tilde{\alpha} + \tilde{\beta}) + \Psi(\sum_{k=1}^{K} \tilde{\alpha}_k)\}} \frac{\exp\{\Psi(\tilde{\lambda}) + \Psi(\tilde{\eta}) + \Psi(\tilde{\lambda}_{kv})\}}{\exp\{2\Psi(\tilde{\lambda} + \tilde{\eta}) + \Psi(\sum_{v=1}^{V} \tilde{\lambda}_{kv})\}} \tag{293}$$

$$\mathscr{L}'(\gamma_{n(K+1)}) = (\Psi(\tilde{\beta}) - \Psi(\tilde{\alpha} + \tilde{\beta}) - \log \gamma_{n(K+1)} - 1) + \lambda$$

From $\mathscr{L}'(\gamma_{n(K+1)}) = 0$, it leads to:
$\gamma_{n(K+1)} \propto \exp\{\mathbb{E}_q[\log(\theta_{d(K+1)})]\}$ or

$$\gamma_{n(K+1)} \propto \exp\{\Psi(\tilde{\beta}) - \Psi(\tilde{\alpha} + \tilde{\beta})\} \tag{294}$$

We obtain two equations for the per-word topic assignment multinomial parameter (variational). We summarize (293) and (294) to form:

$$\begin{cases} \gamma_{nk} \propto \frac{\exp\{\Psi(\tilde{\alpha}_k)+\Psi(\tilde{\alpha})\}}{\exp\{\Psi(\tilde{\alpha}+\tilde{\beta})+\Psi(\sum_{k=1}^{K}\tilde{\alpha}_k)\}} \frac{\exp\{\Psi(\tilde{\lambda})+\Psi(\tilde{\eta})+\Psi(\tilde{\lambda}_{kv})\}}{\exp\{2\Psi(\tilde{\lambda}+\tilde{\eta})+\Psi(\sum_{v=1}^{V}\tilde{\lambda}_{kv})\}}, k \in \{1, 2, 3, ..., K\} \\ \\ \gamma_{n(K+1)} \propto \exp\{\Psi(\tilde{\beta}) - \Psi(\tilde{\alpha} + \tilde{\beta})\}, k = K + 1 \end{cases} \tag{295}$$

The update in (295) is also equivalent to:

$$\begin{cases} \gamma_{nk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \varphi_{kv}] + \mathbb{E}_q[\log(\varphi_{k(V+1)})]\} \\ \\ \gamma_{n(K+1)} \propto \exp\{\mathbb{E}_q[\log(\theta_{d(K+1)})]\} \end{cases} \tag{296}$$

$$\mathscr{L}(\tilde{\lambda}_{kv}) = \sum_{v=1}^{V}(\lambda_{kv} - \tilde{\lambda}_{kv})\mathbb{E}_q[\log \varphi_{kv}] + \sum_{v=1}^{V}\log\Gamma(\tilde{\lambda}_{kv}) - \log\Gamma(\sum_{v=1}^{V}\tilde{\lambda}_{kv}) + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\mathbb{E}_q[\log \varphi_{kv}]$$

$$\mathscr{L}'(\tilde{\lambda}_{kv}) = \Psi'(\tilde{\lambda}_{kv})\left(\lambda_{kv} - \tilde{\lambda}_{kv} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\right) - \Psi'(\sum_{v=1}^{V}\tilde{\lambda}_{kv})\left(\lambda_{kv} - \tilde{\lambda}_{kv} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\right)$$

Computing $\mathscr{L}'(\tilde{\lambda}_{kv}) = 0$, we get: $\left(\lambda_{kv} - \tilde{\lambda}_{kv} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\right) = 0$ or

$$\tilde{\lambda}_{kv} = \lambda_{kv} + \sum_{n=1}^{N}\gamma_{nk}x_n^v \tag{297}$$

Using the steps from (297), we get:

$$\mathscr{L}'(\tilde{\lambda}) = \Psi'(\tilde{\lambda})\left(\lambda - \tilde{\lambda} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\right) - \Psi'(\tilde{\lambda} + \tilde{\eta})\left(\lambda - \tilde{\lambda} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^v\right)$$

$\mathscr{L}'(\tilde{\lambda}) = 0$

$$\tilde{\lambda} = \lambda + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n \tag{298}$$

$$\mathscr{L}'(\tilde{\eta}) = \Psi'(\tilde{\lambda})\left(\eta - \tilde{\eta} + \sum_{n=1}^{N}\gamma_{nk}x_n^{V+1}\right) - \Psi'(\tilde{\eta} + \tilde{\lambda})\left(\eta - \tilde{\eta} + \sum_{n=1}^{N}\gamma_{nk}x_n^{V+1}\right)$$

With $\mathscr{L}'(\tilde{\eta}) = 0$, we have:

$$\tilde{\eta} = \eta + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}x_n^{V+1} \tag{299}$$

$$\mathscr{L}'(\tilde{\alpha}_k) = \Psi(\tilde{\lambda}_k)\left(\alpha_k - \tilde{\alpha}_k + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\right) - \Psi'\left(\sum_{k=1}^{K}\tilde{\alpha}_k\right)\left(\alpha_k - \tilde{\alpha}_k + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\right) - \log\Gamma\left(\sum_{k=1}^{K}\tilde{\alpha}_k\right)$$

$\mathscr{L}'(\tilde{\alpha}_k) = 0$ gives $(\alpha_k - \tilde{\alpha}_k + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}) = 0$ or

$$\tilde{\alpha}_k = \alpha_k + \sum_{n=1}^{N}\gamma_{nk} \tag{300}$$

$$\mathscr{L}'(\tilde{\alpha}) = \Psi'(\tilde{\alpha})\left(\alpha - \tilde{\alpha} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\right) - \Psi'(\tilde{\alpha} + \tilde{\beta})\left(\alpha - \tilde{\alpha} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\right)$$

$$\mathscr{L}'(\tilde{\beta}) = \Psi(\tilde{\beta})\left(\beta - \tilde{\beta} + \sum_{n=1}^{N}\gamma_{n(K+1)}\right) - \Psi'(\tilde{\alpha} + \tilde{\beta})\left(\beta - \tilde{\beta} + \sum_{n=1}^{N}\gamma_{n(K+1)}\right) \tag{301}$$

Setting $\mathscr{L}'(\tilde{\alpha}) = 0$ and $\mathscr{L}'(\tilde{\beta}) = 0$ we get: $(\alpha - \tilde{\alpha} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}) = 0$ and $\left(\beta - \tilde{\beta} + \sum_{n=1}^{N}\gamma_{n(K+1)}\right) = 0$

$$\tilde{\alpha} = \alpha + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk} \tag{302}$$

$$\tilde{\beta} = \beta + \sum_{n=1}^{N}\gamma_{n(K+1)} \tag{303}$$

Chapter 7

Conclusions and Future Work

In this thesis, using a series of directed acyclic graphical models, we formulated efficient alternatives to the latent Dirichlet allocation (LDA) topic model (with its standard symmetric Dirichlet as conjugate prior to the multinomial). Taking advantage of the collapsed representation as it provides a better lower bound (compared to the standard VB method), we were able to implement a series of CVB update equations using the BL or GD priors as alternative to the Dir in chapters 2 and 3. Our proposed inferences in the collapsed space extended the LDA capabilities as it allows us to handle a variety of challenging applications ranging from text document analysis to computer vision (images and videos). The hybrid generative discriminative in chapter 4 uses in the generative stage topic features that are fed into SVM classifiers in the discriminative stage. This hybrid shows in the generative stage the use of BL and GD simultaneously where the SVM accommodates the topic features using efficient probabilistic kernels for classification. Despite the flexibility of the CVB algorithms they are often difficult to characterize. As most Bayesian posteriors, for complex models, are intractable in general, we propose a point estimate (the mode) that offers a much tractable solution. The MAP hypotheses using point estimates are much easier than full Bayesian analysis that integrates over the entire parameter space. We therefore formulate in chapter 5 the MAP-LBLA using standard EM algorithm and it shows that the update equation is much simpler than the ones in the collapsed with the CVB. Importantly, it also shows an implicit equivalent relationship between the MAP and the collapsed representaions with the CVB especially the zero order approximations (CVB0 and the stochastic CVB0). Compared to the situations in parametric topic models where we initially fix the number of topics in advance, we finally propose in chapter 6 a Bayesian nonparametric framework where the HDP prior is the conjugate prior to the multinomial. We formulate the nonprametric prior using the asymmetric BL as a diffuse base measure to enhance variability and heterogeneity in the topics. Our HDP-LBLA with its much bigger hypothesis space ultimately relaxes the assumption of a fixed number of topics and provides an alternative to model selection. From the series of inferences (VB, CGS, and CVB) with the BL and GD, that have been proposed in this thesis, we also emphasize on stochastic optimizations using natural gradients methods to speed up the learning of good topics in online fashion in large scale applications for tasks such as classification and information retrieval. The scheme improves the time and memory complexities. We characterize the efficiency of these priors in topic correlation framework. We ultimately show that asymmetric priors are much robust compared to symmetric priors. As topic models depend extensively on flexible prior distributions in the Bayesian analysis, we show the flexibility of the BL and GD in our proposed topic models. We use both the predictive likelihoods and perplexities as evaluation metods to assess the robustness of our proposed topic models. We have improved object categorization in terms of inferences using the flexibility of these priors. We also improved information retieval system in text document analysis. These two applications present the ultimate capabilities of enhancing a search engine based on topic models (for instance).

Future work could formulate the CVB inferences with other conjugate priors. We could follow a possibility to implement topic models using non conjugate priors, similar to the logistic normal distributions. We could also propose a series of semi-parametric topic models. Another direction could emphasize more on empirical Bayes framework for hyperparameter estimation. Finally, the expectation propagation (EP) is another deterministic approach for topic models learning that could be deployed.

References

[1] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[2] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 524–531, IEEE, 2005.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[4] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34, AUAI Press, 2009.

[5] D. M. Blei, *Probabilistic models of text and images.* PhD thesis, University of California, Berkeley, 2004.

[6] C. Elkan, "Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution," in *Proceedings of the 23rd international conference on Machine learning*, pp. 289–296, 2006.

[7] N. Bouguila, "Count data modeling and classification using finite mixtures of distributions," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 186–198, 2011.

[8] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[9] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking lda: why priors matter," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 1973–1981, Curran Associates Inc., 2009.

[10] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009* (A. P. Danyluk, L. Bottou, and M. L. Littman, eds.), vol. 382 of *ACM International Conference Proceeding Series*, pp. 1105–1112, ACM, 2009.

[11] N. Bouguila, "Hybrid generative/discriminative approaches for proportional data modeling and classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2184–2202, 2012.

[12] Y. W. Teh, D. Newman, and M. Welling, “A collapsed variational bayesian inference algorithm for latent dirichlet allocation,” in *Advances in neural information processing systems*, pp. 1353–1360, 2007.

[13] J. Foulds, L. Boyles, C. DuBois, P. Smyth, and M. Welling, “Stochastic collapsed variational bayesian inference for latent dirichlet allocation,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 446–454, ACM, 2013.

[14] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical dirichlet processes,” *Machine Learning*, pp. 1–30, 12 2006.

[15] Y. W. Teh and M. I. Jordan, “Hierarchical bayesian nonparametric models with applications,” *Bayesian nonparametrics*, vol. 1, pp. 158–207, 2010.

[16] Y. W. Teh, K. Kurihara, and M. Welling, “Collapsed variational inference for hdp,” in *Advances in neural information processing systems*, pp. 1481–1488, 2008.

[17] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, “Fast online em for big topic modeling,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 675–688, 2015.

[18] K. L. Caballero, J. Barajas, and R. Akella, “The generalized dirichlet distribution in enhanced topic detection,” in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 773–782, ACM, 2012.

[19] W. Li, D. Blei, and A. McCallum, “Nonparametric bayes pachinko allocation,” in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 243–250, AUAI Press, 2007.

[20] W. Li and A. McCallum, “Pachinko allocation: Dag-structured mixture models of topic correlations,” in *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584, ACM, 2006.

[21] D. P. Putthividhya, H. T. Attias, and S. Nagarajan, “Independent factor topic models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 833–840, ACM, 2009.

[22] D. P. Putthividhya, *A family of statistical topic models for text and multimedia documents.* PhD thesis, University of California at San Diego, 2010.

[23] A. S. Bakhtiari and N. Bouguila, “A variational bayes model for count data learning and classification,” *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 176–186, 2014.

[24] R. J. Connor and J. E. Mosimann, “Concepts of independence for proportions with a generalization of the dirichlet distribution,” *Journal of the American Statistical Association*, vol. 64, no. 325, pp. 194–206, 1969.

[25] Y. Rui and T. Huang, “Optimizing learning in image retrieval,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 236–243, IEEE, 2000.

[26] C. B. Akgul, B. Sankur, Y. Yemez, and F. Schmitt, "Similarity learning for 3d object retrieval using relevance feedback and risk minimization," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 392–407, 2010.

[27] B. Hu, Y. Liu, S. Gao, R. Sun, and C. Xian, "Parallel relevance feedback for 3d model retrieval based on fast weighted-center particle swarm optimization," *Pattern Recognition*, vol. 43, no. 8, pp. 2950–2961, 2010.

[28] D. Giorgi, M. Mortara, and M. Spagnuolo, "3d shape retrieval based on best view selection," in *Proceedings of the ACM workshop on 3D object retrieval*, pp. 9–14, ACM, 2010.

[29] G. Leifman, R. Meir, and A. Tal, "Semantic-oriented 3d shape retrieval using relevance feedback," *The Visual Computer*, vol. 21, no. 8-10, pp. 865–875, 2005.

[30] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Advances in neural information processing systems*, pp. 1353–1360, 2007.

[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[32] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 159–168, ACM, 1998.

[33] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, ACM, 1999.

[34] N. Bouguila and D. Ziou, "A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2010.

[35] A. S. Bakhtiari and N. Bouguila, "A latent beta-liouville allocation model," *Expert Systems with Applications*, vol. 45, pp. 260–272, 2016.

[36] N. Bouguila, "Clustering of count data using generalized dirichlet multinomial distributions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 462–474, 2008.

[37] A. C. Damianou, M. K. Titsias, and N. D. Lawrence, "Variational inference for latent variables and uncertain inputs in gaussian processes," *Journal of Machine Learning Research (JMLR)*, vol. 2, 2015.

[38] R. Nallapati, "Discriminative models for information retrieval," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 64–71, ACM, 2004.

[39] D. M. Blei, M. I. Jordan, *et al.*, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[40] I. Sato and H. Nakagawa, "Rethinking collapsed variational bayes inference for lda," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 763–770, Omnipress, 2012.

[41] B. Leng, J. Zeng, M. Yao, and Z. Xiong, "3d object retrieval with multitopic model combining relevance feedback and lda model," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 94–105, 2015.

[42] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th annual international conference on machine learning*, pp. 1105–1112, ACM, 2009.

[43] M. A. Abuzneid and A. Mahmood, "Enhanced human face recognition using lbph descriptor, multi-knn, and back-propagation neural network," *IEEE Access*, vol. 6, pp. 20641–20651, 2018.

[44] T. Bdiri and N. Bouguila, "Bayesian learning of inverted dirichlet mixtures for svm kernels generation," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1443–1458, 2013.

[45] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International journal of computer vision*, vol. 73, no. 2, pp. 213–238, 2007.

[46] C. Hentschel and H. Sack, "Does one size really fit all?: Evaluating classifiers in bag-of-visual-words classification," in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, p. 7, ACM, 2014.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178, 2006.

[49] W. Fan and N. Bouguila, "Face detection and facial expression recognition using a novel variational statistical framework," in *International Conference on Multimedia Communications, Services and Security*, pp. 95–106, Springer, 2012.

[50] W. Fan and N. Bouguila, "Learning finite beta-liouville mixture models via variational bayes for proportional data clustering.," in *IJCAI*, pp. 1323–1329, 2013.

[51] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[52] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[53] S. Vishwakarma and A. Agrawal, “A survey on activity recognition and behavior understanding in video surveillance,” *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[54] I. Laptev and T. Lindeberg, “Velocity adaptation of space-time interest points,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, pp. 52–56, IEEE, 2004.

[55] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 32–36, IEEE, 2004.

[56] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[57] N. Bouguila, D. Ziou, and R. I. Hammoud, “On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling,” *Pattern Anal. Appl.*, vol. 12, no. 2, pp. 151–166, 2009.

[58] N. Bouguila, “On the smoothing of multinomial estimates using liouville mixture models and applications,” *Pattern Analysis and Applications*, vol. 16, no. 3, pp. 349–363, 2013.

[59] W. Fan and N. Bouguila, “Online facial expression recognition based on finite beta-liouville mixture models,” in *Computer and Robot Vision (CRV), 2013 International Conference on*, pp. 37–44, IEEE, 2013.

[60] W. Fan and N. Bouguila, “Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference,” *IEEE transactions on neural networks and learning systems*, vol. 24, no. 11, pp. 1850–1862, 2013.

[61] W. Fan and N. Bouguila, “Model-based clustering based on variational learning of hierarchical infinite beta-liouville mixture models,” *Neural Processing Letters*, vol. 44, no. 2, pp. 431–449, 2016.

[62] W. Fan and N. Bouguila, “Online data clustering using variational learning of a hierarchical dirichlet process mixture of dirichlet distributions,” in *International Conference on Database Systems for Advanced Applications*, pp. 18–32, Springer, 2014.

[63] Y. Gao, J. Chen, and J. Zhu, “Streaming gibbs sampling for lda model,” *arXiv preprint arXiv:1601.01142*, 2016.

[64] A. S. Bakhtiari and N. Bouguila, “Online learning for two novel latent topic models,” pp. 286–295, 2014.

[65] M. Hoffman, F. R. Bach, and D. M. Blei, “Online learning for latent dirichlet allocation,” in *advances in neural information processing systems*, pp. 856–864, 2010.

[66] O. Cappé and E. Moulines, “On-line expectation–maximization algorithm for latent data models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[67] K. E. Ihou and N. Bouguila, “A new latent generalized dirichlet allocation model for image classification,” in *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*, pp. 1–6, IEEE, 2017.

[68] D. Mimno, M. Hoffman, and D. Blei, “Sparse stochastic inference for latent dirichlet allocation,” *arXiv preprint arXiv:1206.6425*, 2012.

[69] N. Rasiwasia and N. Vasconcelos, “Latent dirichlet allocation models for image classification,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 2665–2679, 2013.

[70] J. Foulds, L. Boyles, C. DuBois, P. Smyth, and M. Welling, “Stochastic collapsed variational bayesian inference for latent dirichlet allocation,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 446–454, ACM, 2013.

[71] D. G. Lowe, “Object recognition from local scale-invariant features,” *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, 1999.

[72] X. Feng, M. Pietikäinen, and A. Hadid, “Facial expression recognition based on local binary patterns,” *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007.

[73] M. N. Dailey, C. Joyce, M. J. Lyons, M. Kamachi, H. Ishi, J. Gyoba, and G. W. Cottrell, “Evidence and a computational explanation of cultural differences in facial expression recognition.,” *Emotion*, vol. 10, no. 6, p. 874, 2010.

[74] M. J. Lyons, J. Budynek, and S. Akamatsu, “Automatic classification of single facial images,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[75] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 200–205, IEEE, 1998.

[76] F. Cheng, J. Yu, and H. Xiong, “Facial expression recognition in jaffe dataset based on gaussian process classification,” *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1685–1690, 2010.

[77] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[78] T. Kanade, Y. Tian, and J. F. Cohn, “Comprehensive database for facial expression analysis,” in *fg*, p. 46, IEEE, 2000.

[79] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101, IEEE, 2010.

[80] K. E. Ihou and N. Bouguila, “Variational-based latent generalized dirichlet allocation model in the collapsed space and applications,” vol. 332, pp. 372–395, Elsevier, 2019.

[81] A. D. Holub, M. Welling, and P. Perona, “Combining generative models and fisher kernels for object recognition,” in *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, vol. 1, pp. 136–143, IEEE, 2005.

[82] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Advances in neural information processing systems*, pp. 841–848, 2002.

[83] S. Ullman, M. Vidal-Naquet, and E. Sali, “Visual features of intermediate complexity and their use in classification,” *Nature neuroscience*, vol. 5, no. 7, p. 682, 2002.

[84] M. Weber, M. Welling, and P. Perona, “Towards automatic discovery of object categories,” in *cvpr*, vol. 2, p. 39, 2000.

[85] R. Fergus, P. Perona, A. Zisserman, *et al.*, “Object class recognition by unsupervised scale-invariant learning,” in *CVPR (2)*, pp. 264–271, 2003.

[86] B. Leibe and B. Schiele, “Scale-invariant object categorization using a scale-adaptive mean-shift search,” in *Joint Pattern Recognition Symposium*, pp. 145–153, Springer, 2004.

[87] H. Schneiderman, “Learning a restricted bayesian network for object detection,” *CVPR (2)*, vol. 4, pp. 639–646, 2004.

[88] L. Fei-Fei, “Learning generative visual models from few training examples,” in *Workshop on Generative-Model Based Vision, IEEE Proc. CVPR*, 2004.

[89] C. M. Bishop, *Pattern recognition and machine learning.* Springer Science+ Business Media, 2006.

[90] C. Yeh, Y. H. Tsai, and Y. F. Wang, “Generative-discriminative variational model for visual recognition,” *CoRR*, vol. abs/1706.02295, 2017.

[91] W. Roth, R. Peharz, S. Tschiatschek, and F. Pernkopf, “Hybrid generative-discriminative training of gaussian mixture models,” *Pattern recognition letters*, vol. 112, pp. 131–137, 2018.

[92] W. Zheng, Y. Liu, H. Lu, and H. Tang, “Discriminative topic sparse representation for text categorization,” in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, pp. 454–457, IEEE, 2017.

[93] T. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in *Advances in neural information processing systems*, pp. 487–493, 1999.

[94] T. Jebara, R. Kondor, and A. Howard, “Probability product kernels,” *Journal of Machine Learning Research*, vol. 5, no. Jul, pp. 819–844, 2004.

[95] N. Vasconcelos, P. Ho, and P. Moreno, “The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition,” in *European Conference on Computer Vision*, pp. 430–441, Springer, 2004.

[96] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller, "A new discriminative kernel from probabilistic models," in *Advances in Neural Information Processing Systems*, pp. 977–984, 2002.

[97] K. R. Prasad, M. Mohammed, and R. Noorullah, "Visual topic models for healthcare data clustering," *Evolutionary Intelligence*, pp. 1–17, 2019.

[98] L. Xia, D. Luo, C. Zhang, and Z. Wu, "A survey of topic models in text classification," in *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 244–250, IEEE, 2019.

[99] H. J. Steinhauer, T. Helldin, G. Mathiason, and A. Karlsson, "Topic modeling for anomaly detection in telecommunication networks," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2019.

[100] L. Laib, M. S. Allili, and S. Ait-Aoudia, "A probabilistic topic model for event-based image classification and multi-label annotation," *Signal Processing: Image Communication*, vol. 76, pp. 283–294, 2019.

[101] F. Yao and Y. Wang, "Tracking urban geo-topics based on dynamic topic model," *Computers, Environment and Urban Systems*, p. 101419, 2019.

[102] R. Venkatesaramani, D. Downey, B. Malin, and Y. Vorobeychik, "A semantic cover approach for topic modeling," in *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, (Minneapolis, Minnesota), pp. 92–102, Association for Computational Linguistics, June 2019.

[103] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 306–312, AAAI Press, 2014.

[104] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen, "Cqarank: jointly model topics and expertise in community question answering," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 99–108, ACM, 2013.

[105] B. Ghorbani, H. Javadi, and A. Montanari, "An instability in variational inference for topic models," in *International Conference on Machine Learning*, pp. 2221–2231, 2019.

[106] A. Y. Zhang and H. H. Zhou, "Theoretical and computational guarantees of mean field variational inference for community detection," *arXiv preprint arXiv:1710.11268*, 2017.

[107] K. E. Ihou and N. Bouguila, "Stochastic topic models for large scale and nonstationary data," *Eng. Appl. Artif. Intell.*, vol. 88, 2020.

[108] P. Bhagat and P. Choudhary, "Image annotation: Then and now," *Image and Vision Computing*, vol. 80, pp. 1–23, 2018.

[109] D. Tian and Z. Shi, "A two-stage hybrid probabilistic topic model for refining image annotation," *International Journal of Machine Learning and Cybernetics*, pp. 1–15, 2019.

[110] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, pp. 1385–1392, 2004.

[111] H. Zhao, L. Du, W. Buntine, and G. Liu, "Metalda: a topic model that efficiently incorporates meta information," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 635–644, IEEE, 2017.

[112] P. Kherwa and P. Bansal, "Topic modeling: A comprehensive review," *ICST Transactions on Scalable Information Systems*, p. 159623, 07 2018.

[113] L. Liu, H. Huang, Y. Gao, Y. Zhang, and X. Wei, "Neural variational correlated topic modeling," in *The World Wide Web Conference*, pp. 1142–1152, ACM, 2019.

[114] G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang, "A correlated topic model using word embeddings.," in *IJCAI*, pp. 4207–4213, 2017.

[115] I. Korshunova, H. Xiong, M. Fedoryszak, and L. Theis, "Discriminative topic modeling with logistic lda," in *Advances in Neural Information Processing Systems 32*, pp. 6767–6777, Curran Associates, Inc., 2019.

[116] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, pp. 121–128, 2008.

[117] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 248–256, Association for Computational Linguistics, 2009.

[118] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Advances in neural information processing systems*, pp. 897–904, 2009.

[119] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "The dynamic embedded topic model," *CoRR*, vol. abs/1907.05545, 2019.

[120] R. Chi, B. Wu, and L. Wang, "Expert identification based on dynamic lda topic model," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pp. 881–888, IEEE, 2018.

[121] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, ACM, 2006.

[122] J. Chen, J. Zhu, J. Lu, and S. Liu, "Scalable training of hierarchical topic models," *Proceedings of the VLDB Endowment*, vol. 11, no. 7, pp. 826–839, 2018.

[123] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," *Journal of Machine Learning Research*, vol. 6, no. Sep, pp. 1345–1382, 2005.

[124] Y. Li, C. Liu, M. Zhao, R. Li, H. Xiao, K. Wang, and J. Zhang, "Multi-topic tracking model for dynamic social network," *Physica A: Statistical Mechanics and its Applications*, vol. 454, pp. 51–65, 2016.

[125] I. Espinoza, M. Mendoza, P. Ortega, D. Rivera, and F. Weiss, "Viscovery: Trend tracking in opinion forums based on dynamic topic models," *CoRR*, vol. abs/1805.00457, 2018.

[126] Y. He, C. Lin, W. Gao, and K.-F. Wong, "Dynamic joint sentiment-topic model," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 1, p. 6, 2013.

[127] J. Fenglei, G. Cuiyun, *et al.*, "An online topic modeling framework with topics automatically labeled," in *Proceedings of the 2019 Workshop on Widening NLP*, pp. 73–76, 2019.

[128] C. Gao, J. Zeng, M. R. Lyu, and I. King, "Online app review analysis for identifying emerging issues," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, pp. 48–58, IEEE, 2018.

[129] X. Bui, T. Vu, and K. Than, "Stochastic bounds for inference in topic models," in *International Conference on Advances in Information and Communication Technology*, pp. 582–592, Springer, 2016.

[130] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *2008 eighth IEEE international conference on data mining*, pp. 3–12, IEEE, 2008.

[131] S. Padó and M. Lapata, "Dependency-based construction of semantic space models," *Computational Linguistics*, vol. 33, no. 2, pp. 161–199, 2007.

[132] D. Valdez, A. C. Pickett, and P. Goodson, "Topic modeling: Latent semantic analysis for the social sciences," *Social Science Quarterly*, vol. 99, no. 5, pp. 1665–1679, 2018.

[133] J. Chang and D. Blei, "Relational topic models for document networks," in *Artificial Intelligence and Statistics*, pp. 81–88, 2009.

[134] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian, "Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span," *Bmc Bioinformatics*, vol. 7, no. 1, p. 250, 2006.

[135] S. Xiong, K. Wang, D. Ji, and B. Wang, "A short text sentiment-topic model for product reviews," *Neurocomputing*, vol. 297, pp. 94–102, 2018.

[136] M. Hajjem and C. Latiri, "Combining ir and lda topic modeling for filtering microblogs," *Procedia Computer Science*, vol. 112, pp. 761–770, 2017.

[137] M. Fritz and B. Schiele, "Decomposition, discovery and detection of visual categories using topic models," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[138] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 370–377, IEEE, 2005.

[139] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," 2005.

[140] N. Bouguila, D. Ziou, and J. Vaillancourt, “Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application,” *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1533–1543, 2004.

[141] L. Wu, L. Shen, and Z. Li, “A kernel method based on topic model for very high spatial resolution (vhsr) remote sensing image classification,” *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B7, pp. 399–403, 06 2016.

[142] M. Lienou, H. Maitre, and M. Datcu, “Semantic annotation of satellite images using latent dirichlet allocation,” *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 28–32, 2009.

[143] K. Rematas, M. Fritz, and T. Tuytelaars, “Kernel density topic models: Visual topics without visual words,” in *NIPS workshops, Modern Nonparametric Methods in Machine Learning*, 2012.

[144] V. Nguyen, D. Phung, and S. Venkatesh, “Topic model kernel classification with probabilistically reduced features,” *Journal of Data Science*, vol. 13, no. 2, pp. 323–340, 2015.

[145] P. Hennig, D. Stern, R. Herbrich, and T. Graepel, “Kernel topic models,” in *Artificial Intelligence and Statistics*, pp. 511–519, 2012.

[146] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf, “Learning from distributions via support measure machines,” in *Advances in neural information processing systems*, pp. 10–18, 2012.

[147] Y. Yoshikawa, T. Iwata, and H. Sawada, “Latent support measure machines for bag-of-words data classification,” in *Advances in Neural Information Processing Systems*, pp. 1961–1969, 2014.

[148] K. Than and T. Doan, “Guaranteed inference in topic models,” *arXiv preprint arXiv:1512.03308*, 2015.

[149] A. B. Chan, N. Vasconcelos, and P. J. Moreno, “A family of probabilistic kernels based on information divergence,” *Univ. California, San Diego, CA, Tech. Rep. SVCL-TR-2004-1*, 2004.

[150] J. Lin, “Divergence measures based on the shannon entropy,” *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[151] T. Jebara and R. Kondor, “Bhattacharyya and expected likelihood kernels,” in *Learning theory and kernel machines*, pp. 57–71, Springer, 2003.

[152] R. Kondor and T. Jebara, “A kernel between sets of vectors,” in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 361–368, 2003.

[153] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, “Fast online em for big topic modeling,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 675–688, 2015.

[154] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Advances in neural information processing systems*, pp. 649–657, 2015.

[155] J. Z. Wang, J. Li, and G. Wiederhold, “Simplicity: Semantics-sensitive integrated matching for picture libraries,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 9, pp. 947–963, 2001.

[156] S. Yang and H. Zhang, “Text mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis,” *International Journal of Computer and Information Engineering*, vol. 12, no. 7, pp. 525–529, 2018.

[157] Y. Papanikolaou, J. R. Foulds, T. N. Rubin, and G. Tsoumakas, “Dense distributions from sparse samples: Improved gibbs sampling parameter estimators for LDA,” *J. Mach. Learn. Res.*, vol. 18, pp. 62:1–62:58, 2017.

[158] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[159] T. P. Minka and J. D. Lafferty, “Expectation-propogation for the generative aspect model,” *CoRR*, vol. abs/1301.0588, 2013.

[160] L. Yao, D. Mimno, and A. McCallum, “Efficient methods for topic model inference on streaming document collections,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 937–946, 2009.

[161] H. Robbins, S. Monro, *et al.*, “A stochastic approximation method,” *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[162] S. Burkhardt and S. Kramer, “Online sparse collapsed hybrid variational-gibbs algorithm for hierarchical dirichlet process topic models,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 189–204, Springer, 2017.

[163] K. E. Ihou and N. Bouguila, “A smoothed latent generalized dirichlet allocation model in the collapsed space,” pp. 877–880, 2018.

[164] S. M. Katz, “Distribution of content words and phrases in text and language modelling,” *Natural language engineering*, vol. 2, no. 1, pp. 15–59, 1996.

[165] K. W. Church and W. A. Gale, “Poisson mixtures,” *Natural Language Engineering*, vol. 1, no. 2, pp. 163–190, 1995.

[166] W. Li and A. McCallum, “Pachinko allocation: Scalable mixture models of topic correlations,” *J. of Machine Learning Research. Submitted*, 2008.

[167] C. Wang, J. Paisley, and D. Blei, “Online variational inference for the hierarchical dirichlet process,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.

[168] O. Cappé and E. Moulines, “On-line expectation–maximization algorithm for latent data models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[169] A. Ahmed, “On tight approximate inference of logistic-normal admixture model,” in *In Proceedings of the Eleventh International Conference on Artifical Intelligence and Statistics. Omnipress*, Citeseer, 2007.

[170] I. Sato, K. Kurihara, and H. Nakagawa, “Practical collapsed variational bayes inference for hierarchical dirichlet process,” in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 105–113, 2012.

[171] A. Bleier, “Practical collapsed stochastic variational inference for the hdp,” *arXiv preprint arXiv:1312.0412*, 2013.

[172] T. Furuya and R. Ohbuchi, “Dense sampling and fast encoding for 3d model retrieval using bag-of-visual features,” p. 26, 2009.

[173] S. Burkhardt and S. Kramer, “Online sparse collapsed hybrid variational-gibbs algorithm for hierarchical dirichlet process topic models,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 189–204, Springer, 2017.

[174] T. S. Ferguson, “A bayesian analysis of some nonparametric problems,” *The annals of statistics*, pp. 209–230, 1973.

[175] T. S. Ferguson, “Prior distributions on spaces of probability measures,” *The Annals of Statistics*, pp. 615–629, 1974.

[176] J. Pitman and M. Yor, “The two-parameter poisson-dirichlet distribution derived from a stable subordinator,” *The Annals of Probability*, pp. 855–900, 1997.

[177] Y. W. Teh, “A bayesian interpretation of interpolated kneser-ney,” tech. rep., 2006.

[178] Y. W. Teh, “A hierarchical bayesian language model based on pitman-yor processes,” in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 985–992, 2006.

[179] E. B. Sudderth and M. I. Jordan, “Shared segmentation of natural scenes using dependent pitman-yor processes,” in *Advances in neural information processing systems*, pp. 1585–1592, 2009.

[180] W. L. Buntine and S. Mishra, “Experiments with non-parametric topic models,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 881–890, 2014.

[181] J. Paisley, C. Wang, and D. Blei, “The discrete infinite logistic normal distribution for mixed-membership modeling,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 74–82, 2011.

[182] C. Zhang, C. Ek, X. Gratal, F. Pokorny, and H. Kjellstrom, “Supervised hierarchical dirichlet processes with variational inference,” in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 254–261, 2013.

[183] J. Pitman, “Poisson-dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition,” *Comb. Probab. Comput.*, vol. 11, no. 5, pp. 501–514, 2002.

[184] A. K. McCallum, “Mallet: A machine learning for language toolkit,” *http://mallet. cs. umass. edu*, 2002.

[185] C. Wang and D. M. Blei, “Truncation-free online variational inference for bayesian nonparametric models,” in *Advances in neural information processing systems*, pp. 413–421, 2012.

[186] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” *Journal of the ACM (JACM)*, vol. 57, no. 2, pp. 1–30, 2010.

[187] C. Wang and D. M. Blei, “Variational inference for the nested chinese restaurant process,” in *Advances in Neural Information Processing Systems*, pp. 1990–1998, 2009.

[188] Z. Ghahramani and T. L. Griffiths, “Infinite latent feature models and the indian buffet process,” in *Advances in neural information processing systems*, pp. 475–482, 2006.

[189] Y. W. Teh, D. Grür, and Z. Ghahramani, “Stick-breaking construction for the indian buffet process,” in *Artificial Intelligence and Statistics*, pp. 556–563, 2007.

[190] C. Chen, L. Du, and W. Buntine, “Sampling table configurations for the hierarchical poisson-dirichlet process,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 296–311, Springer, 2011.

[191] X. Li, J. OuYang, and X. Zhou, “Sparse hybrid variational-gibbs algorithm for latent dirichlet allocation,” in *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 729–737, SIAM, 2016.

[192] I. Sato and H. Nakagawa, “Topic models with power-law using pitman-yor process,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 673–682, 2010.

[193] J. Sethuraman, “A constructive definition of dirichlet priors,” *Statistica sinica*, pp. 639–650, 1994.

[194] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, “The infinite pcfg using hierarchical dirichlet processes,” in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 688–697, 2007.