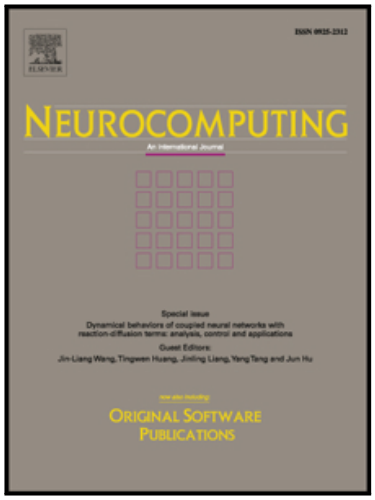
Communicated by Dr Qinfeng Shi

Accepted Manuscript



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

Koffi Eddy Ihou, Nizar Bouguila

PII: S0925-2312(18)31503-0
DOI: https://doi.org/10.1016/j.neucom.2018.12.046
Reference: NEUCOM 20266

To appear in: *Neurocomputing*

Received date: 7 May 2018
Revised date: 12 October 2018
Accepted date: 8 December 2018

Please cite this article as: Koffi Eddy Ihou, Nizar Bouguila, Variational-Based Latent Generalized Dirichlet Allocation Model in the Collapsed Space and Applications, *Neurocomputing* (2018), doi: https://doi.org/10.1016/j.neucom.2018.12.046

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

Koffi Eddy Ihou, Nizar Bouguila*

Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada

Abstract

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.

Keywords: Topic model, generalized Dirichlet, topic correlation, 3D objects, images categorization, facial expression recognition, action recognition in videos.

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 [1] 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) [1, 2, 3]. Topic modeling techniques have been used in a variety of applications, and ultimately led to several extensions of the LDA model. Facing

*I am corresponding author

Email addresses: k_ihou@encs.concordia.ca (Koffi Eddy Ihou), nizar.bouguila@concordia.ca (Nizar Bouguila)

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 [4]. 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 inference. That was the case of models such as CTM (Correlated Topic Models), PAM (Pachinko Allocation Model) [5, 6, 7], IFTM (Independant Factor Topic Models) [8, 9], GD-LDA (Generalized Dirichlet-based LDA)[3], and LGDA (Latent Generalized Dirichlet Allocation) [10]. The GD-LDA for instance is an extension of the original LDA [1] 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 [5, 8]. Despite its success in topic correlation analysis, it leads to a model that is very complex and difficult to implement [5]. Consequently, in the other schemes, the introduction of the GD [11] 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 [12, 13, 14, 15, 16] (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 [1, 4]. 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 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 paper 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 paper is to compare our new approach to the LDA and its previous extensions 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, 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 [1], CVB-LDA [17], GD-LDA [3], and LGDA [10]: as large collection of data creates a large vocabulary size which often leads to a serious sparsity problem, this paper proposes a better prior (GD) that ultimately replaces the traditional Dirichlet distribution. 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 [10]. 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 new 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).

This paper is structured as follows: section 2 illustrates the background and relative work. Section 3 presents the new approach while section 4 covers the experiments and results in several applications. Finally, section 5 explores some future work and provides a conclusion.

2. Related Work And Background

LDA [1] is a generative probabilistic model that has been introduced to solve problems in the original pLSI (probablistic Latent Semantic Indexing) [18, 19, 20]: overfitting and the difficulty in predicting documents probability outside the training set [4]. Known as a multinomial PCA (Principal Component Analysis), the LDA has especially found today its applications in text modeling and computer vision [17]. 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 hyper-parameters 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 paper and for modeling purpose, the variables w and x could be used interchangeably to denote a codeword in an image, 3D object, and video while the variable $\mathcal{X}$ defines a collection of x codewords within the BoW framework.

2.1. Differences in the generative process

Despite our approach being compared to the CVB-LDA [17], GD-LDA [3], and the LGDA [10], all these topic models follow the same generative and Bayesian hierarchical architecture of the original LDA [1, 17]. Nevertheless, each has a different generative process. Following the generative step defined above for the LDA, we can observe that in GD-LDA model [3], 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 [10], the documents parameters were also drawn from a GD distribution. However, the corpus parameter was not generated; in other words, the step 2 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 [1, 3]. 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 paper and its generative model are described in Algorithm 1 while the full comparison between the previous techniques and our model is provided by Table.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:

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

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

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

2.2. Differences in inference techniques

Before going into details in section 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 [17]. The LGDA is based on the 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 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 [17]. 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

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

Fig.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 [17].

2.3. Previous work on image classification using topic model

The LDA (latent Dirichlet allocation) 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 because it has a very limited covariance structure. Despite the fact that the model in [2] provides a better way to label topics (intermediate representations) using unsupervised learning with the LDA , the authors quickly suggested that the 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 generating richer features in order to be successful in the categorization scheme using topics. Some works have been devoted to find new priors for the LDA model [21, 22]. 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 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 guaranties convergence. So, the method is efficient, but it 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 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 in [23, 3] was to marginalize out the parameters leaving only the latent variables that could be now assumed independent given these parameters. Thus, these works 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 implemented in this paper has a graphical architecture that seems to be similar to the bayesian hierarchical model proposed in [2]. However, there is a major difference between these two 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 paper 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. Briefly classic learning approach and algorithm could be summarized through these following concepts. Given observed variables y and unobserved or latent variables x and the model parameters θ we are maximizing the loglikelihood with respect to θ such that:

$$\pounds(\theta) = \log p(y|\theta) = \log \int p(x, y|\theta)dx \tag{3}$$

Often, the difference between the loglikelihood and the bound is expressed as:

$$\pounds(\theta) - \mathcal{F}(q, \theta) = \log p(y|\theta) - \int q(x) \log \frac{p(x, y|\theta)}{q(x)} dx \tag{4}$$

$$= \log p(y|\theta) - \int q(x) \log \frac{p(x|y, \theta)p(y|\theta)}{q(x)} dx \tag{5}$$

$$= -\int \log \frac{p(x|y, \theta)}{q(x)} dx \tag{6}$$

$$= KL(q(x), p(x|y, \theta) \tag{7}$$

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

$$\pounds(\theta^{k-1}) \underbrace{=}_{E-step} \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \underbrace{\leq}_{M-step} \mathcal{F}(q^{(k)}, \theta^{(k)}) \underbrace{\leq}_{Jensen\ \ inequality} \pounds(\theta^{(k)}) \tag{8}$$

where EM converges to a local optimum of $\pounds$. The variational Bayes EM is shown in Algorithm 2 while ours (MCMC) is illustrated by Algorithm 3 which mainly show the differences in the two models.

Algorithm 2 Variational Bayes Expectation-Maximization (EM)

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

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 the approach in [2] and our method.

3. The New Approach

3.1. Overview

In this paper, 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

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.*51
      Update N^t_{kv}, N^t_k, N^t_{dk}
    end for
  end for
Output: Parameters $\tilde{\theta}_{jks}$ and $\tilde{\varphi}_{kws}$ *using Eq.*52 *and* 53
end procedure

[10, 24] 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.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 paper, the documents are drawn from a class set c. The variables ε and ζ are the

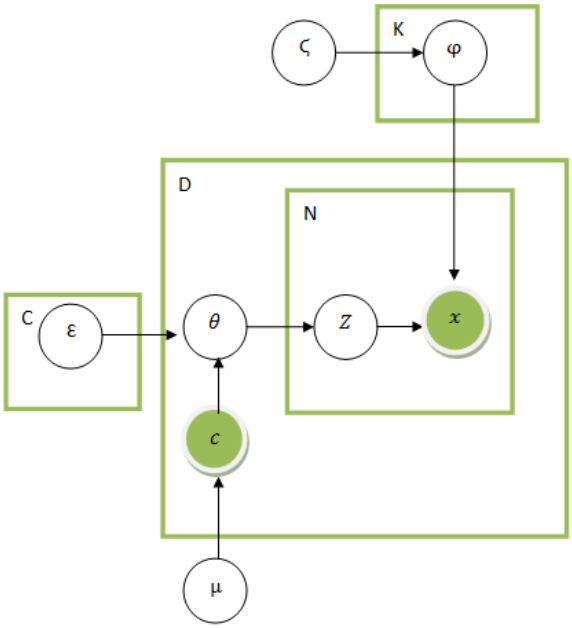


Figure 1: Topic Graphical Model for Classification. The shaded circle denotes observed variables $\boldsymbol{x}$ and the class c.

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 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 [17], 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 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 [25]. This element of the integration functional is also called the variational free energy [17, 23]. 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.1). This concept is similar to [17].

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

3.2. Proposed topic model

In this paper, 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.55) 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 $p(c|\mathcal{X}, \varepsilon, \zeta, \mu)$ [26]. 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 paper 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 a topic model with K as total number of Topics, and N the number of unique words in the dataset, and V as vocabulary size, we can observe that the LDA and the CVB-LDA have similar time complexity, $O(NK)$. The GD-LDA also has the same complexity. While the CVB-LDA and the LDA could only generate topics, the GD-LDA could with the same complexity, perform two tasks: semantic relationship between codewords, and topic correlation analysis. Same time complexity is observed by the LGDA model. In our model,

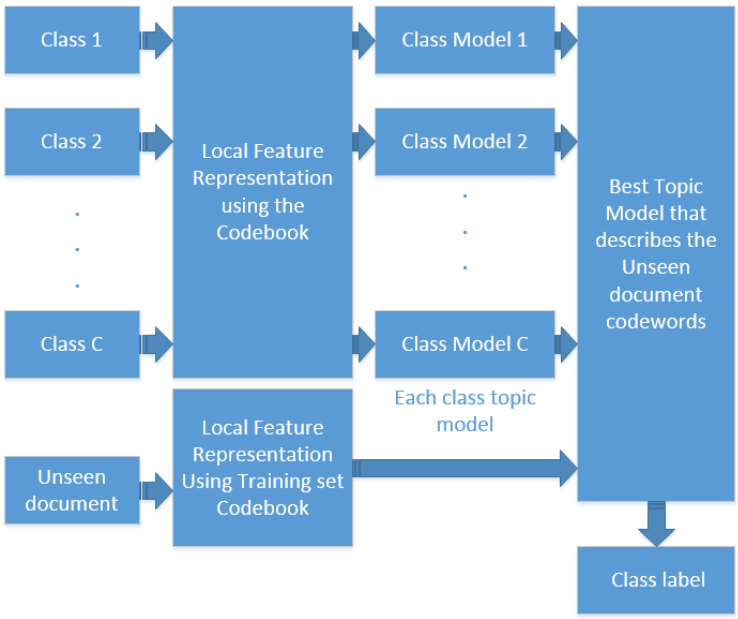


Figure 2: Topic model for classification problem

the CVB-LGDA emphasizes on topic correlation, semantic relationship between words, and codebook analysis bringing his overall time complexity to $O(NKV)$:

$$\left.\begin{array}{rl} for\ n & = 1:N \\ for\ k & = 1:K \\ for\ v & = 1:V \end{array}\right\} \rightarrow Complexity = O(NKV)$$

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, the $O(NKV)$ could be reduced to $O(N)$:

$$\left.\begin{array}{rl} K & \ll N \\ V & \ll N \end{array}\right\} \rightarrow Complexity = O(N) \tag{9}$$

In addition, the variational Bayes-based method and 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 gets its parameters marginalized out leaving only the latent variables. We finally conclude that 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 complexities are summarized in Table 2.

	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	$O(KND)$	No topic correlation as K refers to the K-nearest neighbors (not topics), and D 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: Complexity of the new CVB-LGDA model and other schemes within the BoW framework.

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.

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, 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 [17]. From the graphical model in Fig.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:

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) \prod_{i=1}^{K} p(\varphi_i|\zeta) \prod_{j=1}^{D} p(\theta_j|\varepsilon, c) \times \prod_{n=1}^{N} p(z_{j,n}|\theta_j) p(x_{j,n}|\varphi z_{j,n}) \tag{10}$$

This joint distribution's equation can be simplified to :

$$p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu) = p(c|\mu) p(\theta|c, \varepsilon) p(\varphi|\zeta) \times \prod_{n=1}^{N} p(z_n|\theta) p(x_n|z_n, \varphi) \tag{11}$$

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.

$$p(z, \theta, \varphi|\mathcal{X}, c, \varepsilon, \zeta, \mu) = \frac{p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta, \mu)}{p(\mathcal{X}, c|\varepsilon, \zeta, \mu)} \tag{12}$$

where the denominator is expressed as :

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

with

$$p(\mathcal{X}, c|\varepsilon, \zeta, \mu) = p(\mathcal{X}|\varepsilon, \zeta, c)p(c|\mu) \tag{14}$$

For a uniform class prior, we obtain $p(c|\mu) = p(c) = \frac{1}{C}$ with μ negligible. As a result, the Eq.13 and Eq.14 could be simplified as :

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

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 :

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

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 [17] (factorized distributions) $\tilde{Q}(z, \theta, \varphi)$ such that:

$$\tilde{Q}(z, \theta, \varphi) = \prod_{ij} \tilde{Q}(z_{ij}|\tilde{\psi_{ij}}) \prod_{j} \tilde{Q}(\theta_j|\tilde{\varepsilon_j}) \prod_{k} \tilde{Q}(\varphi_k|\tilde{\zeta_k}) \tag{17}$$

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 [17] that utilizes variational free energy as shown in Eq.18 and Eq.19. 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 Eq.18 to Eq.20, the bound on the loglikelihood is expressed as :

$$\begin{aligned} \log p(\mathcal{X}|c, \varepsilon, \zeta) &\geq \int_{\theta} \int_{\varphi} \sum_{z} Q(z, \theta, \varphi) \times \log p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta) d\varphi d\theta \\ &- \int_{\theta} \int_{\varphi} \sum_{z} Q(z, \theta, \varphi) \log Q(z, \theta, \varphi) d\varphi d\theta \\ &= E_Q[\log p(\mathcal{X}, z, \theta, \varphi, c|\varepsilon, \zeta)] - E_Q[\log Q(z, \theta, \varphi)] \end{aligned} \tag{18}$$

$$
\begin{aligned}
-\log p(\mathcal{X}|c,\varepsilon,\zeta) &\leq -\int_{\theta}\int_{\varphi}\sum_{z} Q(z,\theta,\varphi) \times \log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)d\varphi d\theta \\
&\quad + \int_{\theta}\int_{\varphi}\sum_{z} Q(z,\theta,\varphi)\log Q(z,\theta,\varphi)d\varphi d\theta \\
&= E_{Q}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - E_{Q}[-\log Q(z,\theta,\varphi)]
\end{aligned} \tag{19}
$$

$$
-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq \tilde{\mathcal{F}}(\tilde{Q}(z,\theta,\varphi)) = E_{\tilde{Q}}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - \mathcal{H}(\tilde{Q}(z,\theta,\varphi)) \tag{20}
$$

As the variational entropy is expressed as $\mathcal{H}(\tilde{Q}(z,\theta,\varphi)) = E_{\tilde{Q}}[-\log \tilde{Q}(z,\theta,\varphi)]$, the variational posterior distribution in the joint space $\tilde{Q}(z,\theta,\varphi)$ is factorized using the independency assumption as shown in Eq.17. 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 [1, 2, 17, 27, 10], 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 [17, 28, 29] 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 :

$$
p(\mathcal{X},z,c|\varepsilon,\zeta) = \int_{\theta}\int_{\varphi} p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)d\varphi d\theta \tag{21}
$$

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

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

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 $p(z|\mathcal{X},c,\varepsilon,\zeta)$.

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 [17, 3, 30] 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, is needed. 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 [17]. 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 [17]. This conditional probability is expressed as :

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) = \frac{p(z_{ij}, z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)}{p(z^{-ij}, \mathcal{X}, c, |\varepsilon, \zeta)} \tag{23}$$

The above equation using [17] can be simplified since:

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

The obtained Callen equations (below) as in [17] illustrate the way the collapsed Gibbs actually performs the sampling mechanism. It is an expectation problem as shown in the equation given as:

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

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 [17]. 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.26).

$$\begin{aligned} p(\mathcal{X}, z|c, \varepsilon, \zeta) = C \prod_{j=1}^{D} \left[\prod_{i=1}^{K} \frac{\Gamma(\alpha_{ci} + \beta_{ci})}{\Gamma(\alpha_{ci})\Gamma(\beta_{ci})} \prod_{i=1}^{K} \frac{\Gamma(\alpha'_{ci})\Gamma(\beta'_{ci})}{\Gamma(\alpha'_{ci} + \beta'_{ci})} \right] \\ \times \prod_{j=1}^{D} \left[\prod_{i=1}^{K} \frac{\Gamma(\lambda_r + \eta_r)}{\Gamma(\lambda_r)\Gamma(\eta_r)} \prod_{i=1}^{K} \frac{\Gamma(\lambda'_r)\Gamma(\eta'_r)}{\Gamma(\lambda'_r + \eta'_r)} \right] \end{aligned} \tag{26}$$

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

$$\begin{cases} \alpha'_{ci} = \alpha_{ci} + N^{i}_{j(.)} \\ \beta'_{ci} = \beta_{ci} + \sum_{l=i+1}^{K+1} N^{l}_{j(.)} \end{cases} \tag{27}$$

The topic-word update is defined as :

$$\begin{cases} \lambda'_r = \lambda_r + N^i_{(.),r} \\ \eta'_r = \eta_r + \sum_{d=v+1}^{V+1} N^i_{(.)d} \end{cases} \tag{28}$$

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 (Eq.27, 28, and 29)

$$\begin{cases} N^i_{j(.)} = N^{ij}_{jk(.)} = N^{ij}_{jk.} \\ N^l_{j(.)} = N^{ij}_{jl(.)} = N^{ij}_{jl.} \\ N^i_{(.),r} = N^{ij}_{(.),k\nu_{ij}} = N^{ij}_{.k\nu_{ij}} \\ N^i_{(.)d} = N^{ij}_{(.),kd} = N^{ij}_{.kd} \end{cases} \tag{29}$$

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.30, 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.30 is obtained in a collapsed space, 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:

$$\hat{\psi}_{ijkc} = p(z_{ij} = k|\mathcal{X}, c, \varepsilon, \zeta) \tag{30}$$

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.23 so that:

$$\begin{aligned} p(z_{ij} = k|z^{-ij}, \mathcal{X}, c, \varepsilon, \zeta) \propto & \left[\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N^{-ij}_{jl.})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N^{-ij}_{jl.})} \right] \\ & \times \left[\frac{(N^{-ij}_{.k\nu_{ij}} + \lambda_\nu)(\eta_\nu + \sum_{d=\nu+1}^{V+1} N^{-ij}_{.kd_{ij}})}{(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} N^{-ij}_{.kd_{ij}})} \right] \end{aligned} \tag{31}$$

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

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta) = \frac{A(k)}{B(k', K)} \tag{32}$$

such that :

$$A(k) = \left[\frac{(N_{jk.}^{-ij} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N_{jl.}^{-ij})} \times \frac{(N_{.k\nu_{ij}}^{-ij} + \lambda_{\nu})(\eta_{\nu} + \sum_{d=\nu+1}^{V+1} N_{.kd_{ij}}^{-ij})}{(\lambda_{\nu} + \eta_{\nu} + \sum_{d=\nu}^{V+1} N_{.kd_{ij}}^{-ij})} \right] \tag{33}$$

and

$$B(k', K) = \sum_{k'=1}^{K} \left[\frac{(N_{jk'.}^{-ij} + \alpha_{ck'})(\beta_{ck'} + \sum_{l=k'+1}^{K+1} N_{jl.}^{-ij})}{(\alpha_{ck'} + \beta_{ck'} + \sum_{l=k'}^{K+1} N_{jl.}^{-ij})} \frac{(N_{.k'\nu_{ij}}^{-ij} + \lambda_{\nu})(\eta_{\nu} + \sum_{d=\nu+1}^{V+1} N_{.k'd_{ij}}^{-ij})}{(\lambda_{\nu} + \eta_{\nu} + \sum_{d=\nu}^{V+1} N_{.k'd_{ij}}^{-ij})} \right] \tag{34}$$

Now, the collapsed Gibbs sampler uses the Callen equations (Eq.25) as in [17] 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 Eq.30. 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,

$$\tilde{\psi}_{ijkc} \neq \hat{\psi}_{ijkc} \tag{35}$$

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.17 from Eq.20 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 [17].

$$\tilde{\alpha}_{jkc} = \alpha_c + \sum_i \tilde{\psi}_{ijkc} \tag{36}$$

$$\tilde{\beta}_{jk'c} = \beta_c + \sum_i \tilde{\psi}_{ijk'c} \tag{37}$$

$$\tilde{\lambda}_{kw} = \lambda + \sum_{ij} \vec{1}(x_{ij} = w)\tilde{\psi}_{ijkc} \tag{38}$$

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

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 [17].
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 [17, 1, 4], 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 [17]. 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.

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 [17], 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:

$$\hat{Q}(z, \theta, \varphi) = \hat{Q}(\theta, \varphi|z) \prod_{ij} \hat{Q}(z_{ij}|\hat{\psi}_{ij}) \tag{40}$$

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:

$$\hat{\mathcal{F}}(\hat{Q}(z)\hat{Q}(\theta,\varphi|z)) = E_{\hat{Q}(z)\hat{Q}(\theta,\varphi|z)}[-\log p(\mathcal{X},z,\theta,\varphi,c|\varepsilon,\zeta)] - \mathcal{H}(\hat{Q}(z)\hat{Q}(\theta,\varphi|z)) \tag{41}$$

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

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 [17]. 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 :

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

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

$$-\log p(\mathcal{X}|c,\varepsilon,\zeta) \leq \hat{\mathcal{F}}(\hat{Q}(z)) = E_{\hat{Q}(z)}[-\log p(\mathcal{X},z,c|\varepsilon,\zeta)] - \mathcal{H}(\hat{Q}(z)) \tag{44}$$

$$\hat{\mathcal{F}}(\hat{Q}(z)) \leq \tilde{\mathcal{F}}(\tilde{Q}(z)) \triangleq \min_{\tilde{Q}(\theta)\tilde{Q}(\varphi)} \tilde{\mathcal{F}}(\tilde{Q}(z)\tilde{Q}(\theta)\tilde{Q}(\varphi)) \tag{45}$$

Eq.45 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.44 with respect to ψ_{ijk} leads to the multinomial update in each class as shown in Eq.46.

$$\hat{\psi}_{ijkc} = \hat{Q}(z_{ij}=k|c) = \frac{exp(E_{\hat{Q}(z^{-ij})}[p(\mathcal{X},z^{-ij},z_{ij}=k,c|\varepsilon,\zeta])}{\sum_{k'=1}^{K} exp(E_{\hat{Q}(z^{-ij})}[p(\mathcal{X},z^{-ij},z_{ij}=k',c|\varepsilon,\zeta])} \tag{46}$$

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 reached in [17]

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 [17]. 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 [17]. So, the mean of the sum of the Bernoulli variables means and variance of the sum of the Bernoulli variable variances [17] are respectively computed as :

$$E_{\hat{Q}}[N_{jkc.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{i'jkc} \tag{47}$$

$$Var_{\hat{q}}[N_{jkc.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{i'jkc}(\vec{1} - \hat{\psi}_{i'jkc}) \tag{48}$$

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 [17], we obtained:

$$E_{\hat{Q}}[\log(\alpha + N_{jkc.})] \approx \log(\alpha + E_{\hat{Q}}[N_{jkc.}]) - \frac{Var_{\hat{Q}}[N_{jkc.}]}{2(\alpha + E_{\hat{Q}}[N_{jkc.}])^2} \tag{49}$$

Therefore, the expression above becomes:

$$\exp(E_{\hat{Q}}[\log(\alpha + N_{jkc.})]) \approx (\alpha + E_{\hat{Q}}[N_{jkc.}]) - \exp\left(\frac{Var_{\hat{Q}}[N_{jkc.}]}{2(\alpha + E_{\hat{Q}}[N_{jkc.}])^2}\right) \tag{50}$$

This is the second-order Taylor expansion used as an approximation [31]. 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 :

$$\begin{aligned}
\hat{Q}(z_{ij} = k|c) = \hat{\psi}_{ijkc} \propto \\
\left\{ \left[\frac{(\alpha_{ck} + E_{\hat{Q}}[N_{jk.}^{-ij}])(\beta_{ck} + \sum_{l=k+1}^{K+1} E_{\hat{Q}}[N_{jl.}^{-ij}])}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} E_{\hat{Q}}[N_{jl.}^{-ij}])} \right] \right. \\
\times \left[\frac{(\lambda_{\nu} + E_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}])(\eta_{\nu} + \sum_{d=\nu+1}^{V+1} E_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])}{(\lambda_{\nu} + \eta_{\nu} + \sum_{d=\nu}^{V+1} E_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])} \right] \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{jk.}^{-ij})}{2(\alpha_{ck} + E_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{2(\beta_{ck} + (\sum_{l=k+1}^{K+1} E_{\hat{Q}}[N_{jl.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}((N_{.k\nu_{ij}}^{-ij})}{2(\lambda_{\nu} + E_{\hat{Q}}[N_{.k\nu_{ij}}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{l=k+1}^{K+1} N_{jl.}^{-ij})}{2(\alpha_{ck} + \beta_{ck} + E_{\hat{Q}}[\sum_{l=k+1}^{K+1} N_{jl.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{d=\nu+1}^{V+1} N_{.kd_{ij}}^{-ij})}{2(\eta_{\nu} + (\sum_{d=\nu+1}^{V+1} E_{\hat{Q}}[N_{.kd_{ij}}^{-ij}])^2}\right) \\
\left. \times \exp\left(\frac{Var_{\hat{Q}}(\sum_{d=\nu}^{V+1} N_{.kd_{ij}}^{-ij})}{2(\lambda_{\nu} + \eta_{\nu} + (\sum_{d=\nu}^{V+1} E_{\hat{Q}}[n_{.kd_{ij}}^{-ij}])^2}\right) \right\}
\end{aligned} \tag{51}$$

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

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 [17], document parameter distribution is given as:

$$\hat{\theta}_{jk} = \frac{(\alpha_{kc} + E_Q[N_{jk.}])(\beta_{kc} + \sum_{l=k}^{K+1} E_Q[N_{jk.}])}{(\alpha_{kc} + \beta_{kc} + \sum_{l=k}^{K+1} E_Q[N_{jk.}])} \tag{52}$$

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

$$\hat{\varphi}_{kw} = \frac{(\lambda_v + E_Q[N_{.kv_{ij}}])(\eta_v + \sum_{d=v+1}^{V+1} E_Q[N_{.kd_{ij}}])}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} E_Q[N_{.kd_{ij}}])} \tag{53}$$

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 [3]. The goal is to compute efficiently the probability of the held-out dataset [32, 3]. After estimation of the predictive distributions, we used the empirical likelihood estimate scheme presented in [3] as a validation method. In the CVB-LGDA model, the likelihood [17, 3] could be reduced to:

$$p(\mathcal{X}_{unseenDoc}) = p(\mathcal{X}_{unseenDoc}|c, \varepsilon, \zeta) = \prod_{ij} \sum_{k} \hat{\theta}_{jk} \hat{\varphi}_{kw} \tag{54}$$

such that the counts $E_Q[N_{jk.}]$, $E_Q[N_{.kv_{ij}}]$, and $E_Q[N_{.kd_{ij}}]$ 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 [3]. 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 wordd probability distributions 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 [3, 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.

3.5. Bayesian decision boundary for classification

The empirical likelihood estimate provides the probability of seeing the test set. In our classification problem (Fig.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:

$$p(c|\mathcal{X}, \mu, \varepsilon, \zeta) \propto p(\mathcal{X}|c, \varepsilon, \zeta)p(c|\mu) \propto p(\mathcal{X}|c, \varepsilon, \zeta) \tag{55}$$

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:

$$C^* = \underset{c}{\operatorname{argmax}}\, p(\mathcal{X}|c, \varepsilon, \zeta) \tag{56}$$

3.6. Model Selection

It is really challenging in topic modeling framework to choose and fix the number of topics. As already explained in [4], 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.

4. Experimental results

In the topic modeling literature, several applications have often focused on text modeling. In our experiments in this paper, 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 paper mainly emphasize on representations using local features.

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

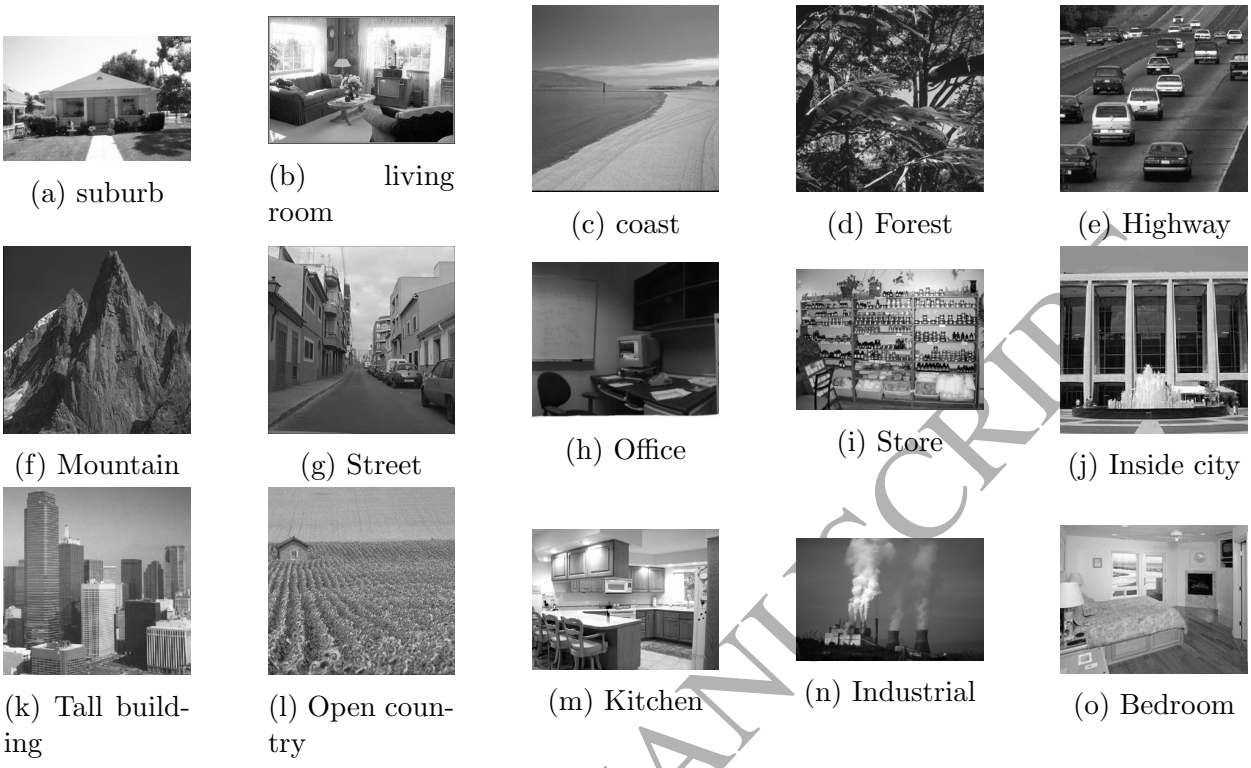
4.1. Image Categorization

4.1.1. Methodology

In our experiments, we constructed our model using the well-known grayscale 15 categories natural scenes dataset [33] . As illustrated in Fig.3 and Table 4, this widely known and challenging data set includes the following 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 Transform) algorithm (Fig.17). 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 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.

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.4). In model selection (Fig.6), the optimal number of topics obtained is $K = 145$ while the optimal vocabulary size is $V = 1450$. The overall accuracy rate is



(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 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 4: size of each image category.

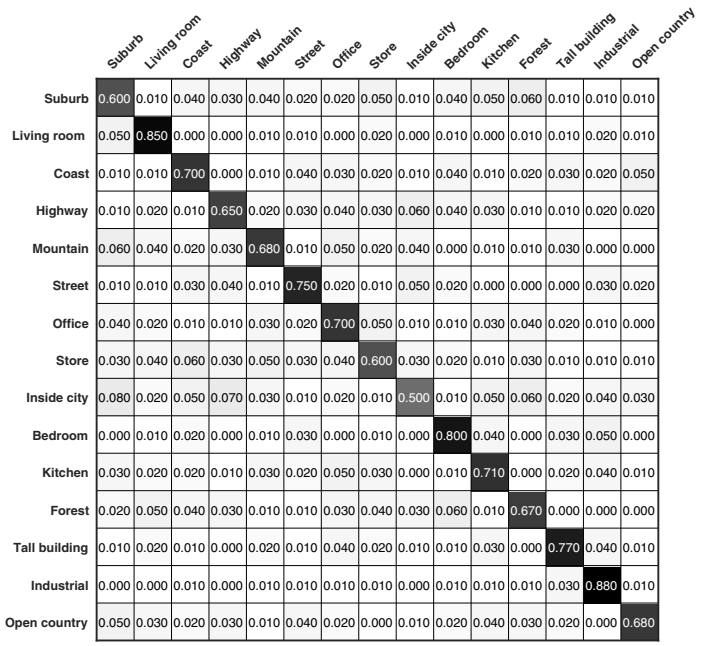


Figure 4: Confusion matrix for the natural scenes classification problem.

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.5) 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.

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 [34, 35]).
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 [36] for facial expression images modeling. We considered the JAFFE (Japanese Female Facial Expression) dataset (See Fig.7 and 8). It contains 213 images obtained from 10 Japanese females showing 7 facial expressions such as surprise, anger, happiness, sadness,

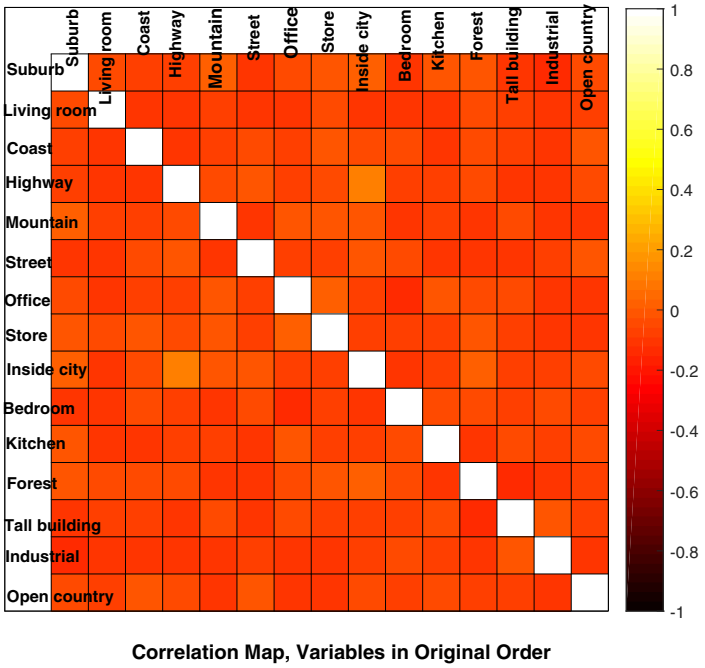


Figure 5: Natural scene images correlation map.

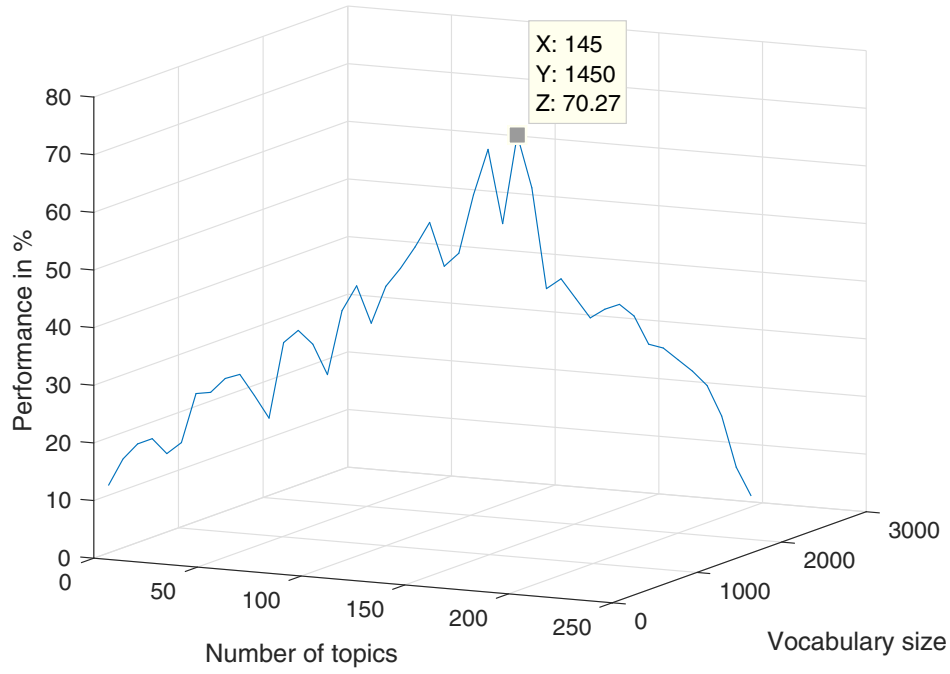


Figure 6: Optimal number of topics and vocabulary size for image classification problem.

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.9). 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.10) obtained shows high accuracy rate of 89.8% as shown in Fig.12. which outperforms its competitors (see Table 3). In addition, the optimal number of topics is $K = 70$ while the optimal vocabulary size is $V = 105$. We illustrated a correlation map (Fig.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.

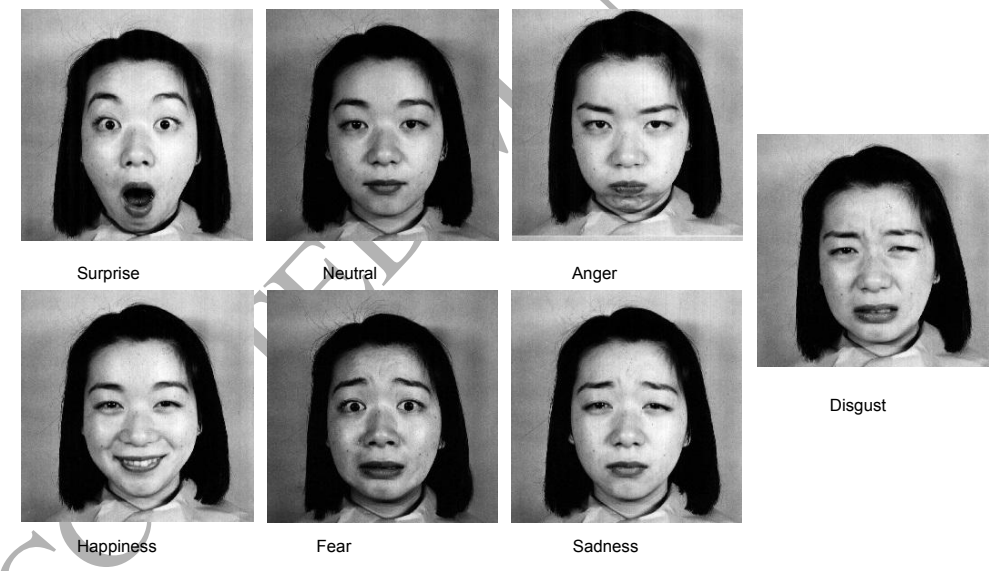


Figure 7: Facial expressions and emotions in the JAFFE dataset

4.3. 3D object classification

The dataset (Fig.18) we consider in this application contains 10 classes of 3D objects [37]. 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.18. For the training set, 7 (3D) objects are randomly selected with

Figure 8: Women showing a "surprised" facial expression

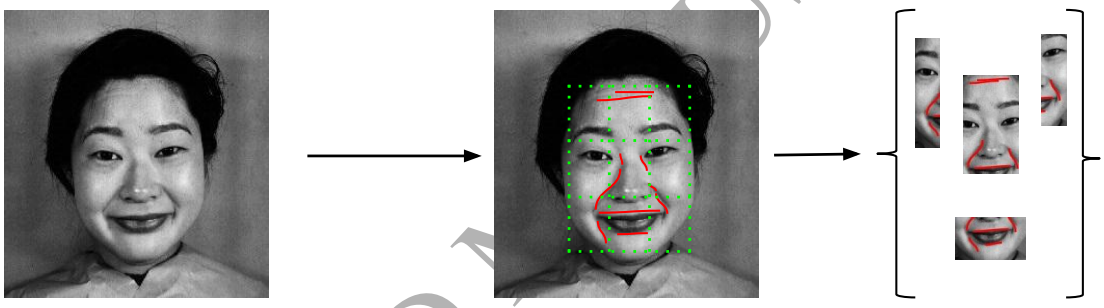


Figure 9: Facial Expression: Key Regions of Interest and Extraction

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 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.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.

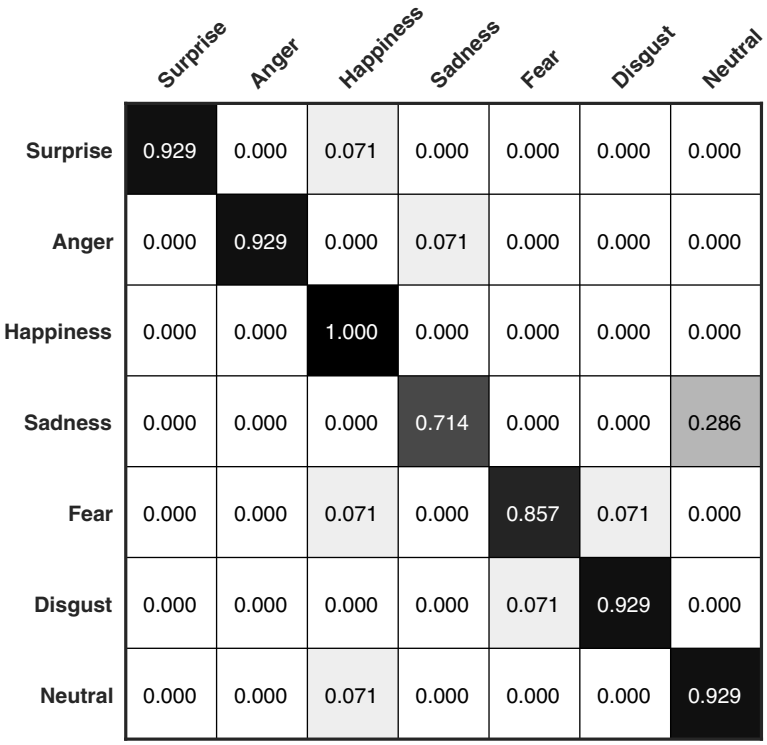


Figure 10: Confusion matrix from the Facial expressions classification

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.15), the accuracy rate shown by the confusion matrix (Fig.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.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. 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.14.

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 [38]. In this paper, 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 [39, 40]. It mainly includes individuals (25 actors) in 4 scenarios performing 6 types of human ac-

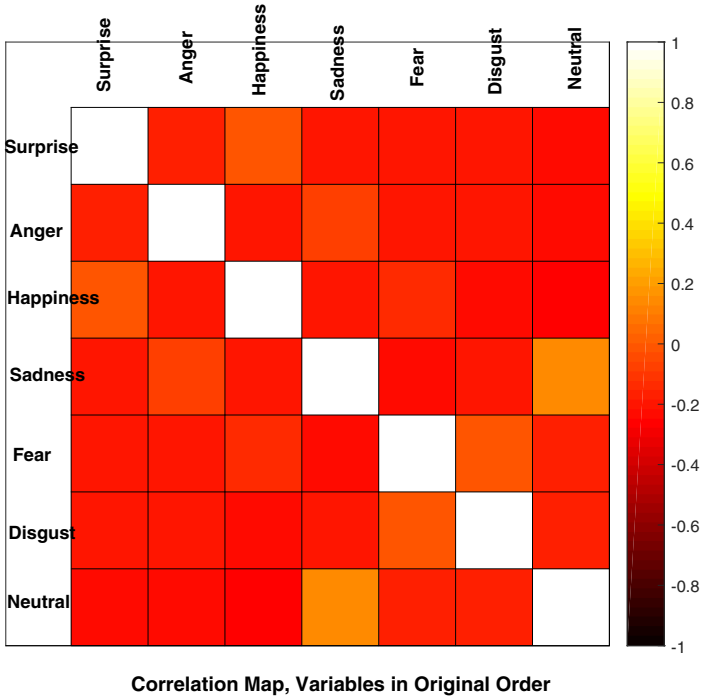


Figure 11: Correlation map for facial expression categories

tions (walking, running, jogging, boxing, hand waving, and hand clapping) as illustrated in Table.9. 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 [41]. 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 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 [22], 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

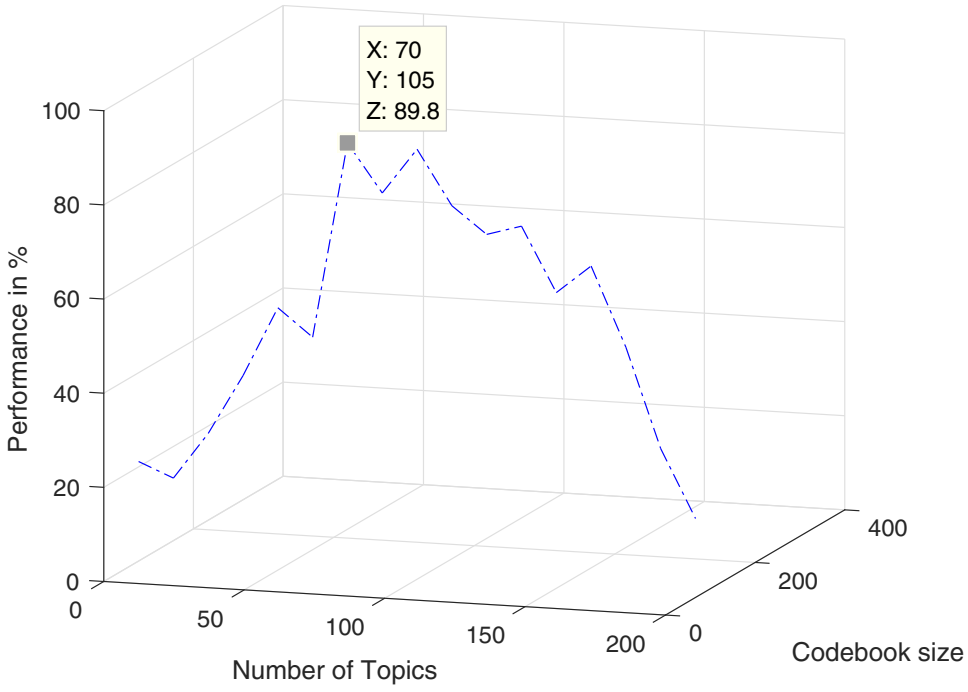


Figure 12: Model selection for facial expressions

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), we can say that the CVB-LGDA model is very robust and could be definitely an alternative to finite mixture models considering its performances [42, 43].

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.

4.5. Classification results with other supervised models

To evaluate our proposed model and inference technique, we set up a goal to compare the new approach with K-Nearest Neighbor (KNN), Backpropagation Neural Network (BPNN), and SVM. In our settings, 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. As our entire collections have a BoW feature representation, the distance of choice in case of the KNN was the Euclidean distance. 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

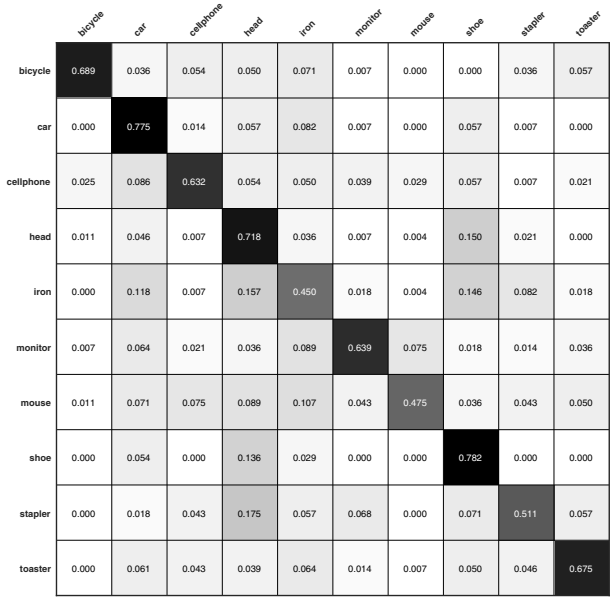


Figure 13: 3D object confusion matrix

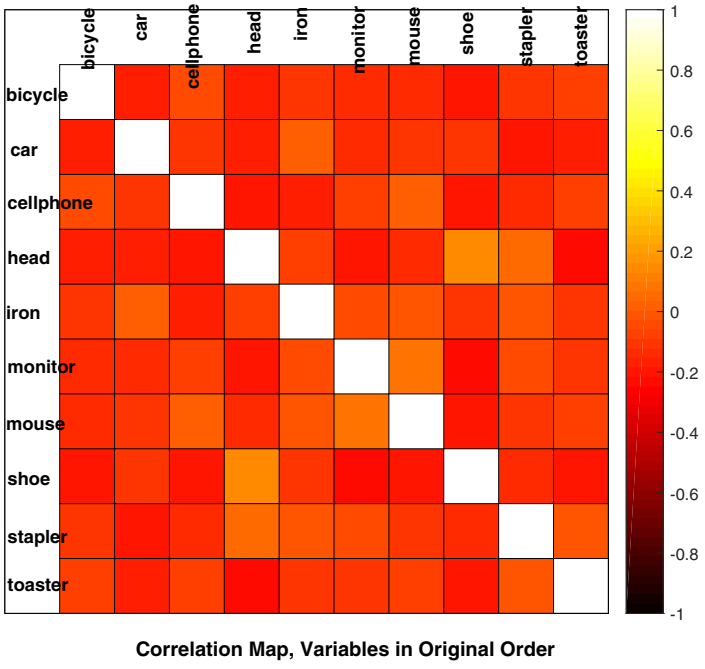


Figure 14: Correlation map for the 3D objects categories.

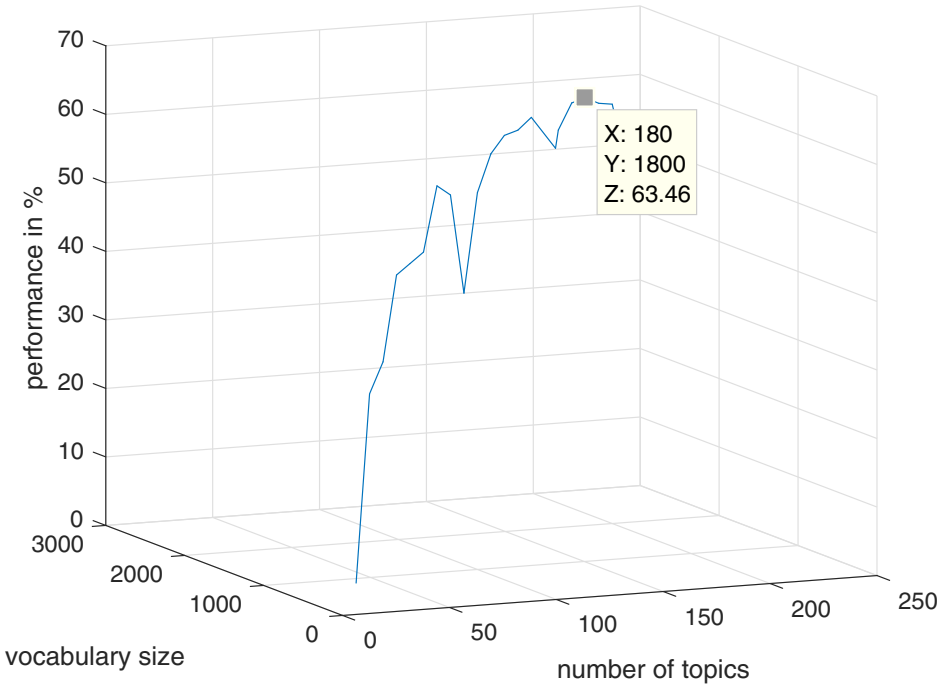


Figure 15: Optimal number of topics and vocabulary size for 3D modeling

Figure 16: 2D Features extraction for a 3D modeling

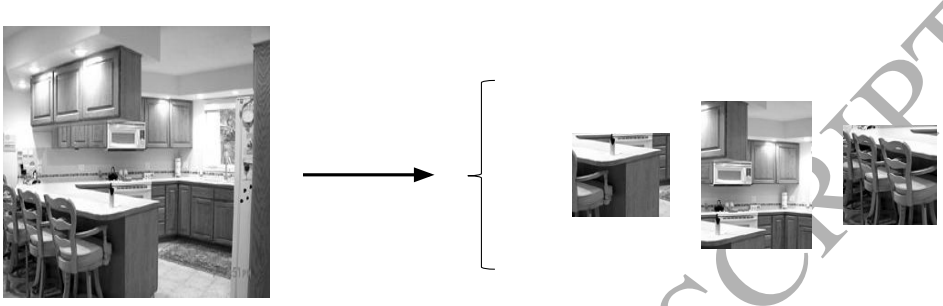


Figure 17: Natural scene image Features extraction

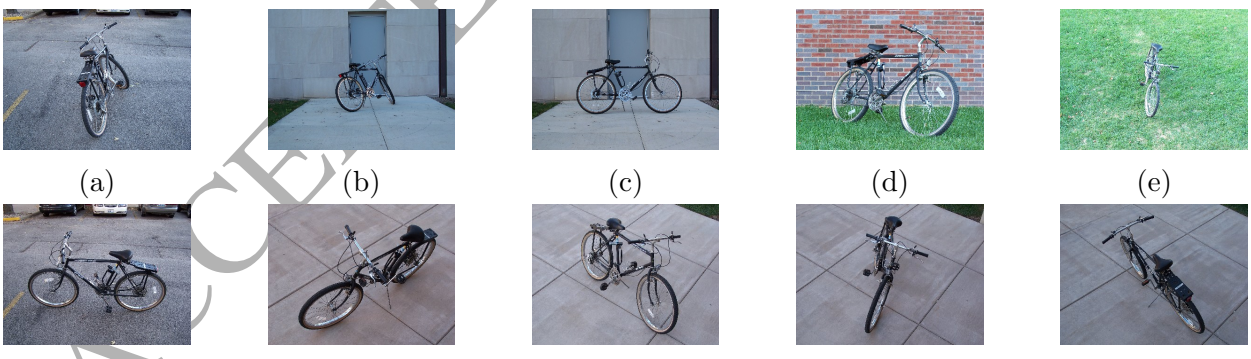


Figure 18: An object from a bicycle's class at different 2D views for a 3D modeling

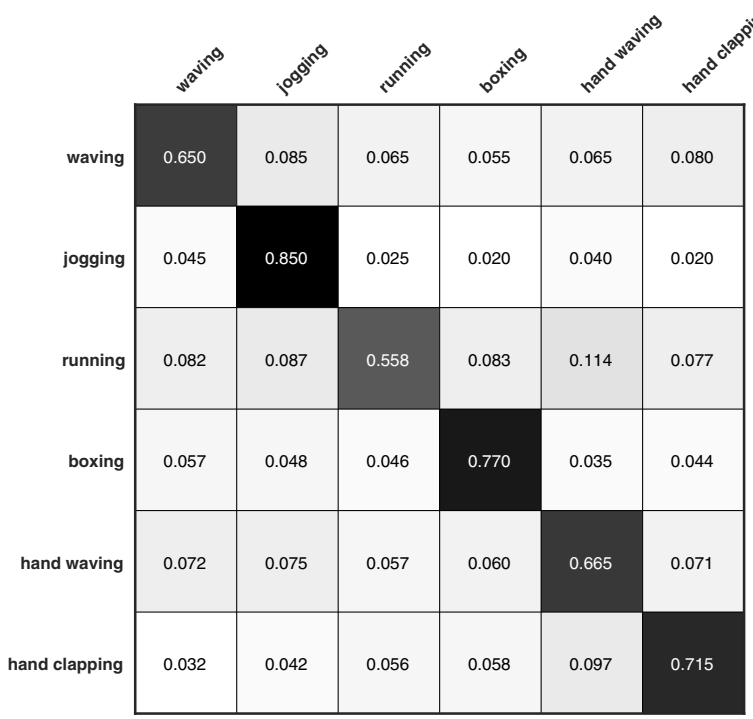


Figure 19: Confusion matrix of the action classes in video

average accuracy values obtained from these datasets are summarized in Tables 5, 6, 7, and 8. From these tables 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 $K = 7$). In addition, KNN provides good performance in the case of low dimensional data than in the case of high dimensional data (videos and 3Ds) due to the large vocabulary size.
In SVM, we considered Radial Basis Function (RBF) Kernel. The kernel parameter (A) is taken from $\{0.1, 1.0, 4\}$. The results in terms of averaged classification accuracy obtained in Tables 5, 6, 7, and 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 following 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 (L) in the output layer. Though, everything is getting slow as we increase the number of neurons. 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 our 3 tested classifiers: 68.4%,

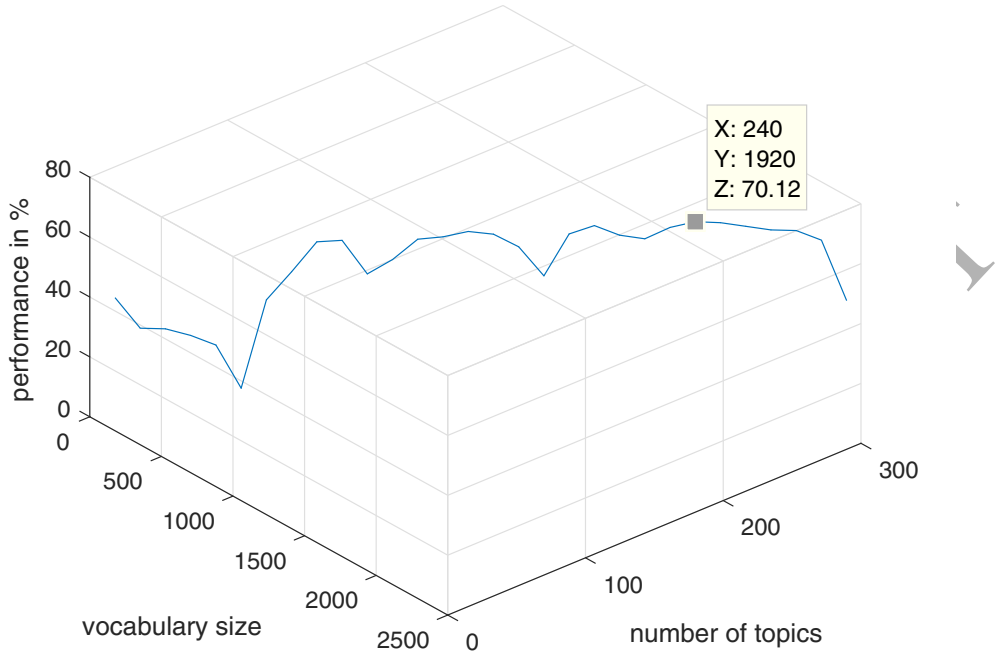


Figure 20: Model selection for actions using videos

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 5: Performance of BPNN, SVM and KNN using images (natural scene).

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.

5. Conclusion

In this paper, 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

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 6: Performance of BPNN, SVM and KNN using face expressions.

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 7: Performance of BPNN, SVM and KNN using 3D objects.

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

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 8: Performance of BPNN, SVM and KNN using action recognition datasets.

boxing	hand clapping	jogging	hand waving	Running	walking

Table 9: KTH Action Recognition Dataset

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 other classification approaches such as KNN, SVM and BPNN. 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. 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. The model could also be improved to be executed in an online fashion to cope with situations where new documents could recursively update the codeword distributions in the database.

Acknowledgements

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

[2] L. Fei-Fei, 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, IEEE, 2005, pp. 524–531.

[3] K. L. Caballero, J. Barajas, R. Akella, The generalized dirichlet distribution in enhanced topic detection, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012, pp. 773–782.

[4] D. M. Blei, Probabilistic models of text and images, Ph.D. thesis, University of California, Berkeley (2004).
[5] D. Blei, J. Lafferty, Correlated topic models, Advances in neural information processing systems 18 (2006) 147.
[6] W. Li, D. Blei, A. McCallum, Nonparametric bayes pachinko allocation, in: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2007, pp. 243–250.
[7] W. Li, A. McCallum, Pachinko allocation: Dag-structured mixture models of topic correlations, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 577–584.
[8] D. P. Putthividhya, H. T. Attias, S. Nagarajan, Independent factor topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 833–840.
[9] D. P. Putthividhya, A family of statistical topic models for text and multimedia documents, Ph.D. thesis, University of California at San Diego (2010).
[10] A. S. Bakhtiari, N. Bouguila, A variational bayes model for count data learning and classification, Engineering Applications of Artificial Intelligence 35 (2014) 176–186.
[11] R. J. Connor, J. E. Mosimann, Concepts of independence for proportions with a generalization of the dirichlet distribution, Journal of the American Statistical Association 64 (325) (1969) 194–206.
[12] Y. Rui, T. Huang, Optimizing learning in image retrieval, in: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, Vol. 1, IEEE, 2000, pp. 236–243.
[13] C. B. Akgul, B. Sankur, Y. Yemez, F. Schmitt, Similarity learning for 3d object retrieval using relevance feedback and risk minimization, International Journal of Computer Vision 89 (2-3) (2010) 392–407.
[14] B. Hu, Y. Liu, S. Gao, R. Sun, C. Xian, Parallel relevance feedback for 3d model retrieval based on fast weighted-center particle swarm optimization, Pattern Recognition 43 (8) (2010) 2950–2961.
[15] D. Giorgi, M. Mortara, M. Spagnuolo, 3d shape retrieval based on best view selection, in: Proceedings of the ACM workshop on 3D object retrieval, ACM, 2010, pp. 9–14.
[16] G. Leifman, R. Meir, A. Tal, Semantic-oriented 3d shape retrieval using relevance feedback, The Visual Computer 21 (8-10) (2005) 865–875.
[17] Y. W. Teh, D. Newman, M. Welling, A collapsed variational bayesian inference algorithm for latent dirichlet allocation, in: Advances in neural information processing systems, 2007, pp. 1353–1360.
[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (6) (1990) 391.
[19] C. H. Papadimitriou, H. Tamaki, P. Raghavan, S. Vempala, Latent semantic indexing: A probabilistic analysis, in: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, ACM, 1998, pp. 159–168.
[20] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.
[21] N. Bouguila, D. Ziou, A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling, IEEE Transactions on Neural Networks 21 (1) (2010) 107–122.
[22] A. S. Bakhtiari, N. Bouguila, A latent beta-liouville allocation model, Expert Systems with Applications 45 (2016) 260–272.
[23] Y. W. Teh, D. Newman, M. Welling, A collapsed variational bayesian inference algorithm for latent dirichlet allocation, in: Advances in neural information processing systems, 2007, pp. 1353–1360.
[24] N. Bouguila, Clustering of count data using generalized dirichlet multinomial distributions, IEEE Transactions on Knowledge and Data Engineering 20 (4) (2008) 462–474.
[25] A. C. Damianou, M. K. Titsias, N. D. Lawrence, Variational inference for latent variables and uncertain inputs in gaussian processes, Journal of Machine Learning Research (JMLR) 2.
[26] R. Nallapati, Discriminative models for information retrieval, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2004, pp. 64–71.
[27] D. M. Blei, M. I. Jordan, et al., Variational inference for dirichlet process mixtures, Bayesian analysis 1 (1) (2006) 121–144.
[28] I. Sato, H. Nakagawa, Rethinking collapsed variational bayes inference for lda, in: Proceedings of the

29th International Coference on International Conference on Machine Learning, Omnipress, 2012, pp. 763–770.

[29] J. Foulds, L. Boyles, C. DuBois, P. Smyth, 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, ACM, 2013, pp. 446–454.

[30] B. Leng, J. Zeng, M. Yao, Z. Xiong, 3d object retrieval with multitopic model combining relevance feedback and lda model, IEEE Transactions on Image Processing 24 (1) (2015) 94–105.

[31] A. Asuncion, M. Welling, P. Smyth, Y. W. Teh, On smoothing and inference for topic models, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 27–34.

[32] H. M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 1105–1112.

[33] S. Lazebnik, C. Schmid, 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, 2006, pp. 2169–2178. doi:10.1109/CVPR.2006.68.

[34] W. Fan, N. Bouguila, Face detection and facial expression recognition using a novel variational statistical framework, in: International Conference on Multimedia Communications, Services and Security, Springer, 2012, pp. 95–106.

[35] W. Fan, N. Bouguila, Learning finite beta-liouville mixture models via variational bayes for proportional data clustering., in: IJCAI, 2013, pp. 1323–1329.

[36] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence 29 (6) (2007) 915–928.

[37] S. Savarese, L. Fei-Fei, 3d generic object categorization, localization and pose estimation, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.

[38] S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, The Visual Computer 29 (10) (2013) 983–1009.

[39] I. Laptev, T. Lindeberg, Velocity adaptation of space-time interest points, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, Vol. 1, IEEE, 2004, pp. 52–56.

[40] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, Vol. 3, IEEE, 2004, pp. 32–36.

[41] B. K. Horn, B. G. Schunck, Determining optical flow, Artificial intelligence 17 (1-3) (1981) 185–203.

[42] N. Bouguila, D. Ziou, R. I. Hammoud, On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling, Pattern Anal. Appl. 12 (2) (2009) 151–166. doi:10.1007/s10044-008-0111-4.

[43] N. Bouguila, Count data modeling and classification using finite mixtures of distributions, IEEE Trans. Neural Networks 22 (2) (2011) 186–198. doi:10.1109/TNN.2010.2091428.

Biography

Koffi Eddy Ihou is currently a PhD student in the Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC. His current research interests include pattern recognition, machine learning, and computer vision.

Nizar Bouguila received the B.E. degree in computer science from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently a Full Professor with the Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC. His current research interests include pattern recognition, machine learning, and computer vision.