Variational Learning for Finite Inverted Dirichlet Mixture Models and Its Applications

Parisa Tirdad

A Thesis

in

The Concordia Institute

for

Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Information Systems Security) at

Concordia University

Montréal, Québec, Canada

March 2015

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By : **Parisa Tirdad**

Entitled : **Variational Learning for Finite Inverted Dirichlet Mixture Models and Its Applications**

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

Master of Applied Science in Information Systems Security

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

Signed by the final examining committee :

____ Chair

Dr. Mohammad Mannan

____ Examiner

Dr. Jamal Bentahar

____ Examiner

Dr. Hoi Dick Ng

____ Supervisor

Dr. Nizar Bouguila

____ Co-supervisor

Dr. Djemel Ziou

Approved by Jamal Bentahar

Chair of Department or Graduate Program Director

Date: March-25-2015. ____

Dr. Amir Asif

Dean of Faculty of Engineering and Computer Science

Abstract

Variational Learning for Finite Inverted Dirichlet Mixture Models and Its Applications

Parisa Tirdad

Clustering is an important step in data mining, machine learning, computer vision and image processing. It is the process of assigning similar objects to the same subset. Among available clustering techniques, finite mixture models have been remarkably used, since they have the ability to consider prior knowledge about the data. Employing mixture models requires, choosing a standard distribution, determining the number of mixture components and estimating the model parameters. Currently, the combination of Gaussian distribution, as the standard distribution, and Expectation Maximization (EM), as the parameter estimator, has been widely used with mixture models. However, each of these choices has its own limitations. In this thesis, these limitations are discussed and addressed via defining a variational inference framework for finite inverted Dirichlet mixture model, which is able to provide a better capability in modeling multivariate positive data, that appear frequently in many real world applications. Finite inverted Dirichlet mixtures enable us to model high-dimensional, both symmetric and asymmetric data. Compared to the conventional expectation maximization (EM) algorithm, the variational approach has the following advantages: it is computationally more efficient, it converges fast, and is able to estimate the parameters and the number of the mixture model components, automatically and simultaneously. The experimental results validate the presented approach on different synthetic datasets and shows its performance for two interesting and challenging real world applications, namely natural scene categorization and human activity classification.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Nizar Bouguila, and my co-supervisor professor Djemel Ziou for their invaluable guidance, comments and criticisms, throughout the course of my master program. I have been privileged to work under their supervision, and I truly appreciate their help.

Special thanks to Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this thesis.

I thank my fellow lab-mates at Concordia University, for the good thoughts, kindness, and motivation during my master's program.

Many thanks to my parents, Parvin and Mohammadreza, my brother Kaveh and my sister Shiva, for all the unconditional love, constant support and understanding.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction and Related Works

Advances in technology has allowed to generate and store large amounts of multimodal data (text, image, video, audio). A crucial problem is the statistical modeling and analysis of these data. This is evidenced by many information retrieval systems [3, 4], and various data mining and machine learning techniques. Clustering, in particular, has been the topic of extensive research in the past and several parametric and nonparametric approaches have been proposed [5–11]. Among these approaches finite mixtures have received a lot of attention. The most popular mixture model is the Gaussian mixture [12–14] which has been applied in several applications [15–17]. However, this choice is not appropriate when the data partitions are not Gaussians as shown in several previous works [1, 18–21]. Indeed, it is important to take the form of the data and the kind of latent structure it expresses into account. This is exactly the case of positive data for which the inverted Dirichlet, that we shall consider in this thesis, has been shown to be an efficient alternative to the Gaussian [1, 21, 22].

Two challenging problems when dealing with finite mixtures are the determination of the number of the mixture components and the estimation of the mixture's parameters. Concerning parameters estimation, two families of approaches could be considered namely the frequentist and the Bayesian techniques. The maximum likelihood (ML), while the most popular among frequentist estimation techniques, to mixture learning has several shortcomings for its application since it can easily get caught in saddle points or local maxima and it depends on the initially set parameters. It was implemented in [1], via an expectation-maximization algorithm [23], in the case of the inverted Dirichlet mixture. To address these drawbacks, Bayesian framework could be adopted. Bayesian learning has several interesting properties. For instance, it allows to incorporate prior knowledge in a natural way, it permits the manipulation of uncertainty consistently, and it does not suffer from over-fitting problems. However, fully Bayesian learning is generally computationally intractable which has forced researchers in the past to adopt approximation techniques such as Laplace approximation and Markov Chain Monte Carlo (MCMC) sampling. In particular MCMC-based sampling approaches have received a lot of attention, yet they suffer from significant computational complexity. Thus, variational learning has been proposed, as an efficient deterministic approximation to fully Bayesian learning, to overcome the problems related to MCMC sampling and have been widely adopted [24]. Variational learning has made it possible to fit large class of learning models and then to explore real-world complexity of data [25, 26]. It can be viewed as an approximation to the exact pure Bayesian learning where the true posterior is approximated with a simpler distribution.

A finite mixture model is the linear combination of finite number of weighted standard distributions (e.g. Gaussian, Dirichlet, inverted Dirichlet) called mixture components, and defined as

$$p(X) = \sum_{j=1}^{M} \pi_j p(X|\theta_j) \tag{1.1}$$

where $p(X|\theta_j)$ is a mixture component with parameter θ_j. Parameters π_j are model weights or mixing coefficients, and $0 \leq \pi_j \leq 1, \sum_{j=1}^{M} \pi_j = 1$. In (1.1), M is considered to be a fixed quantity.

There are some other cases that the number of the components are infinite *i.e.*, $M \to \infty$, these cases are called infinite mixture models and are defined as:

$$p(X) = \sum_{j=1}^{\infty} \pi_j p(X|\theta_j) \tag{1.2}$$

In this thesis, we are focusing on the finite mixture models.

1.2 Contributions

The main contributions of this thesis are as follows:

☞**Proposing a variational framework for finite inverted Dirichlet mixture model** :

Our approach proposes a variational framework for learning inverted Dirichlet mixture models. We build a traceable lower bound for estimating marginal likelihood by using approximated distributions to replace the intractable parameter distributions. This approach is computationally more efficient, and converges fast. Furthermore, in comparison to traditional approaches, in which the model selections are solved based on cross-validation, our method estimates the model parameters and determines the number of components simultaneously.

☞**Demonstrating the application of the proposed statistical model** :

In addition to showing the validity of the proposed approach in parameter estimation and model selection on different synthetic datasets, we have shown the usefulness of our method in challenging real world applications. First application is natural scene categorization which plays an important role in understanding the world through images and information retrieval. Human activity classification is the second application that has attracted lots of attention for its important applications in security systems and surveillance for public environments. Moreover, we compared the performance of our model with Gaussian mixture models in terms of accuracy and showed that the variational inverted Dirichlet model outperforms

Gaussian mixture models in modeling real world data. Finally, to show the importance of choosing a variational approach for estimating the mixture components, we used both variational inference and maximum likelihood approaches with the finite inverted Dirichlet mixture models and compared the results.

1.3 Thesis Overview

The organization of this thesis is as follows:

- ❏ Chapter 1 explains data clustering using mixture models, and the challenges that should be addressed by choosing them. It also gives an overview of the proposed approach.
- ❏ Chapter 2 proposes a variational framework for finite inverted Dirichlet mixture model, which is able to estimate the model parameters and determine the number of components simultaneously.
- ❏ Chapter 3 shows the experimental results of the proposed approach on synthetic data and two real world applications, namely, natural scene categorization and human activity classification.
- ❏ Chapter 4 summarizes the research and presents the conclusions.

Chapter 2

Proposed Statistical Framework

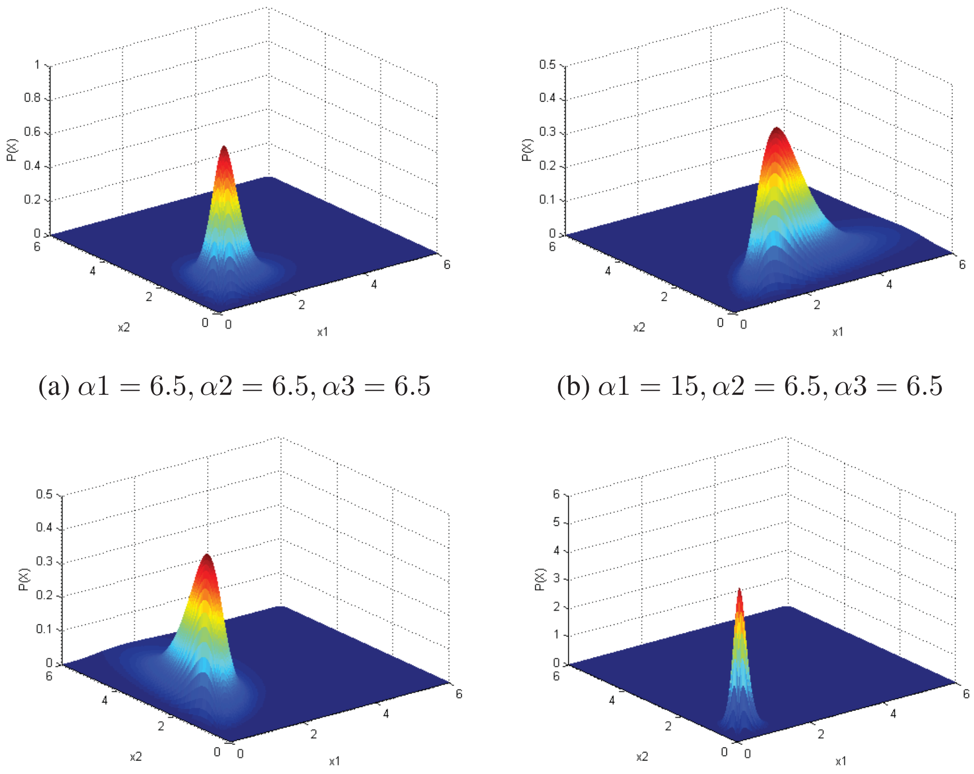
2.1 Introduction

In the previous chapter, we discussed the importance of data clustering and explained some drawbacks of the existing methods. This chapter, gives an overview of the inverted Dirichlet mixture model and defines a variational framework for it, which allows estimating the parameters and the number of components of the mixture model automatically and simultaneously.

2.2 Finite Inverted Dirichlet Mixture Model

The main reason for adopting inverted Dirichlet distribution as the standard distribution for our mixture model is that, inverted Dirichlet can generate models specific to positive data and as shown in Figure (2.1), unlike Gaussian distribution, it is considerably flexible and can perform in both symmetric and asymmetric modes. The inverted Dirichlet distribution has many interesting properties and has applications in various fields [27–29].

Assume that a D-dimensional positive vector $\vec{X}_i = (X_{i1}, \dots, X_{iD})$ is sampled from a finite



(a) $\alpha 1 = 6.5, \alpha 2 = 6.5, \alpha 3 = 6.5$

(b) $\alpha 1 = 15, \alpha 2 = 6.5, \alpha 3 = 6.5$

(c) $\alpha 1 = 6.5, \alpha 2 = 15, \alpha 3 = 6.5$

(d) $\alpha 1 = 6.5, \alpha 2 = 6.5, \alpha 3 = 15$

Figure 2.1: Bivariate inverted Dirichlet distributions, in symmetric and asymmetric modes.

inverted Dirichlet mixture model with M components, then we have:

$$p(\vec{X}_i|\vec{\pi},\vec{\alpha}) = \sum_{j=1}^{M} \pi_j \mathcal{ID}(\vec{X}_i|\vec{\alpha}_j) \tag{2.1}$$

where $\vec{\alpha} = (\vec{\alpha}_1, ..., \vec{\alpha}_M)$ and $\vec{\pi} = (\pi_1, ..., \pi_M)$ denotes the mixing coefficients with the constraints that they are positive and sum to one. $\mathcal{ID}(\vec{X}_i|\vec{\alpha}_j)$ represents the j^{th} inverted Dirichlet distribution with parameter $\vec{\alpha}_j$ and is defined in [27] as

$$\mathcal{ID}(\vec{X}_i|\vec{\alpha}_j) = \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\Pi_{l=1}^{D+1}\Gamma(\alpha_{jl})} \Pi_{l=1}^{D} X_{il}^{\alpha_{jl}-1} (1 + \sum_{l=1}^{D} X_{il})^{-\sum_{l=1}^{D+1} \alpha_{jl}} \tag{2.2}$$

where $0 < X_{il} < \infty$ for $l = 1, ..., D$. In addition, $\vec{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jD})$ such that $\alpha_{j1} > 0$ for $l = 1, ..., D+1$. The mean, variance and covariance of the inverted Dirichlet distribution are given by

$$E(X_l) = \frac{\alpha_l}{(\alpha_{D+1} - 1)} \tag{2.3}$$

$$var(X_l) = \frac{\alpha_l(\alpha_j + \alpha_{D+1} - 1)}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2)} \tag{2.4}$$

$$cov(X_a, X_b) = \frac{\alpha_a \alpha_b}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2)} \tag{2.5}$$

Next, we introduce an M-dimensional binary random vector $\vec{Z}_i = \{Z_{i1}, ..., Z_{iM}\}$ for each observed vector $\vec{X}_i$, such that $Z_{ij} \in \{0, 1\}$, $\sum_{j=1}^{M} Z_{ij} = 1$, and $Z_{ij} = 1$ if $\vec{X}_i$ belongs to component j and 0, otherwise. Notice that, $\mathcal{Z} = \{\vec{Z}_1, ...\vec{Z}_N\}$ are called the *membership vectors* of the mixture model and are also considered as the latent variables since they are actually hidden variables that do not appear explicitly in the model. Furthermore, the conditional distribution of $\mathcal{Z}$ given the mixing coefficients $\vec{\pi}$ is defined as

$$p(\mathcal{Z}|\vec{\pi}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{2.6}$$

Then, the likelihood function with latent variables, which is indeed the conditional distribution of data set $\mathcal{X}$ given the class labels $\mathcal{Z}$ can be written as

$$p(\mathcal{X}|\mathcal{Z}, \vec{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathcal{ID}(\vec{X}_i|\vec{\alpha}_j)^{Z_{ij}} \tag{2.7}$$

Moreover, we assume that the parameters of the inverted Dirichlet are statistically independent and for each parameter α_{jl}, the Gamma distribution $\mathcal{G}$ is adopted to approximate the conjugate prior:

$$p(\alpha_{jl}) = \mathcal{G}(\alpha_{jl}|u_{jl}, v_{jl}) = \frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}^{-1}} e^{-v_{jl}\alpha_{jl}} \tag{2.8}$$

where u_{jl} and v_{jl} are positive hyperparameters. Thus, the joint distribution of all the random variables, conditioned on the mixing coefficients can be written as

$$\begin{aligned} p(\mathcal{X}, \mathcal{Z}, \vec{\alpha}|\vec{\pi}) &= p(\mathcal{X}|\mathcal{Z}, \vec{\alpha}) p(\mathcal{Z}|\vec{\pi}) p(\vec{\alpha}) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{M} \Bigg[\pi_j \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} \prod_{l=1}^{D} X_{il}^{\alpha_{jl}-1} \\ &\quad \times (1 + \sum_{l=1}^{D} X_{il})^{-\sum_{l=1}^{D+1} \alpha_{jl}} \Bigg]^{Z_{ij}} \\ &\quad \times \prod_{j=1}^{M} \prod_{l=1}^{D+1} \frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-v_{jl}\alpha_{jl}} \end{aligned} \tag{2.9}$$

A directed representation of this model is illustrated in Figure (2.2).

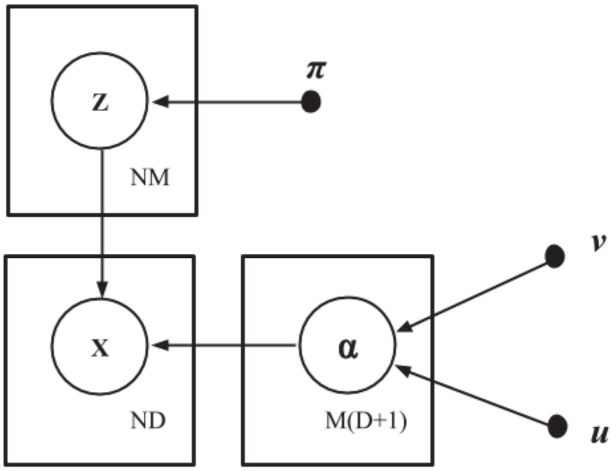


Figure 2.2: Graphical model representation of the finite inverted Dirichlet mixture. Symbols in circles denote random variables; otherwise, they denote model parameters. Plates indicate repetition (with the number of repetitions in the lower right), and arcs describe conditional dependencies between variables.

2.3 Variational Learning

Variational inference is a deterministic approximation scheme, which is used to formulate the computation of a marginal or conditional probability in terms of an optimization problem. In this section, following the methodology proposed in [30], we develop a variational framework for learning the finite inverted Dirichlet mixture model. To simplify the notation without loss of generality, we define $\Theta = \{\mathcal{Z}, \vec{\alpha}\}$. The main idea in variational learning is to find an approximation $Q(\Theta)$, which approximates the true posterior distribution $p(\Theta|\mathcal{X}, \vec{\pi})$. The logarithm of the model evidence $p(\mathcal{X}|\vec{\pi})$ can be decomposed as

$$\ln p(\mathcal{X}|\vec{\pi}) = \mathcal{L}(q) - \underbrace{\int Q(\Theta) \ln \left[\frac{p(\Theta|\mathcal{X}, \vec{\pi})}{Q(\Theta)}\right] d\Theta}_{-KL(Q\|p)} \tag{2.10}$$

where $KL(Q \parallel p)$ is the Kullback-Leibler (KL) divergence between $Q(\Theta)$ and the true posterior distribution $p(\Theta|\mathcal{X}, \vec{\pi})$. $\mathcal{L}(q)$ is the variational lower bound of $\ln p(\mathcal{X})$ and is defined by

$$\mathcal{L}(q) = \int Q(\Theta) \ln \left[\frac{p\left(\mathcal{X}, \Theta | \vec{\pi}\right)}{Q(\Theta)} \right] d\Theta \tag{2.11}$$

In our work, a mean field approximation [31, 32] is adopted for the variational inference. Hence, $Q(\Theta)$ can be factorized into disjoint tractable distributions as follows:

$$Q(\Theta) = Q(\mathcal{Z}) Q(\vec{\alpha}) \tag{2.12}$$

In order to maximize the lower bound $\mathcal{L}(q)$, we need to make a variational optimization of $\mathcal{L}(q)$ with respect to each of the factors in turn. For a specific factor $Q_s(\Theta_s)$ the general variational solution is given by

$$Q_s(\Theta_s) = \frac{\exp \left\langle \ln p\left(\mathcal{X}, \Theta | \vec{\pi}\right) \right\rangle_{\neq s}}{\int \exp \left\langle \ln p\left(\mathcal{X}, \Theta | \vec{\pi}\right) \right\rangle_{\neq s} d\Theta} \tag{2.13}$$

Where $\langle . \rangle_{\neq s}$ denotes an expectation with respect to all factor distributions, except for s. We can obtain the following variational solutions for the finite inverted Dirichlet mixture model (proved in appendix A):

$$Q\left(\mathcal{Z}\right) = \prod_{i=1}^{N} \prod_{j=1}^{M} {r_{ij}}^{Z_{ij}} \tag{2.14}$$

$$Q(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D+1} \mathcal{G}(\alpha_{jl} | u^*_{jl}, v^*_{jl}) \tag{2.15}$$

where we have defined

$$r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M} \rho_{ij}} \tag{2.16}$$

$$\begin{aligned}\rho_{ij} = \exp\Bigg\{ & \ln \pi_j + \tilde{\mathcal{R}}_j + \sum_{l=1}^{D} (\bar{\alpha}_{jl} - 1) \ln X_{il} \\ & - (\sum_{l=1}^{D+1} \bar{\alpha}_{jl}) \ln(1 + \sum_{l=1}^{D} X_{il}) \Bigg\}\end{aligned} \tag{2.17}$$

$$\begin{aligned}\widetilde{\mathcal{R}}_j = & \ln \frac{\Gamma\left(\sum_{l=1}^{D+1} \bar{\alpha}_{jl}\right)}{\prod_{l=1}^{D+1} \Gamma\left(\bar{\alpha}_{jl}\right)} \\ & + \sum_{l=1}^{D+1} \bar{\alpha}_{jl} \left[\Psi\left(\sum_{l=1}^{D+1} \bar{\alpha}_{jl}\right) - \Psi\left(\bar{\alpha}_{jl}\right) \right] \left[\langle \ln \alpha_{jl} \rangle - \ln \bar{\alpha}_{jl} \right] \\ & + \frac{1}{2} \sum_{l=1}^{D+1} \bar{\alpha}_{jl}^2 \left[\Psi'\left(\sum_{l=1}^{D+1} \bar{\alpha}_{jl}\right) - \Psi'\left(\bar{\alpha}_{jl}\right) \right] \left\langle \left(\ln \alpha_{jl} - \ln \bar{\alpha}_{jl}\right)^2 \right\rangle \\ & + \frac{1}{2} \sum_{a=1}^{D+1} \sum_{\substack{b=1 \\ (b \neq a)}}^{D+1} \bar{\alpha}_{ja} \bar{\alpha}_{jb} \Bigg[\Psi'\left(\sum_{l=1}^{D+1} \bar{\alpha}_{jl} \left(\langle \ln \alpha_{ja} \rangle - \ln \bar{\alpha}_{ja} \right) \right) \\ & \times \left(\langle \ln \alpha_{jb} \rangle - \ln \bar{\alpha}_{jb} \right) \Bigg]\end{aligned} \tag{2.18}$$

$$\begin{aligned}u_{jl}^* = u_{jl} + \sum_{i=1}^{N} \langle Z_{ij} \rangle \bar{\alpha}_{jl} \Bigg[& \Psi(\sum_{l=1}^{D+1} \bar{\alpha}_{jl}) - \Psi(\bar{\alpha}_{jl}) \\ & + \sum_{k \neq l}^{D+1} \bar{\alpha}_k \Psi'(\sum_{l=1}^{D+1} \bar{\alpha}_l)(\langle \ln \alpha_k \rangle - \ln \bar{\alpha}_k) \Bigg]\end{aligned} \tag{2.19}$$

$$v_{jl}^* = v_{jl} - \sum_{i=1}^{N} \langle Z_{ij} \rangle \left[\ln X_{il} - \ln \left(1 + \sum_{l=1}^{D} X_{il} \right) \right] \tag{2.20}$$

where $\Psi\left(.\right)$ is diagamma function. The expected values in the above formulas are

$$\langle Z_{ij} \rangle = r_{ij} \tag{2.21}$$

$$\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}}{v_{jl}} \tag{2.22}$$

$$\langle \ln \alpha_{jl} \rangle = \Psi \left(u_{jl} \right) - \ln v_{jl} \tag{2.23}$$

Note that, $\tilde{\mathcal{R}}_j$ is the approximate lower bound of $\mathcal{R}_j$, where $\mathcal{R}_j$ is defined as $\mathcal{R}_j = \left\langle \ln \frac{\Gamma\left(\sum_{l=1}^{D+1} \alpha_{jl}\right)}{\prod_{l=1}^{D+1} \Gamma\left(\alpha_{jl}\right)} \right\rangle$. Since a closed form expression cannot be found for $\mathcal{R}_j$, the standard variational inference can not be applied directly. Therefore, we applied the second-order Taylor series expansion to find a lower bound approximation $\tilde{\mathcal{R}}_j$ for the variational inference (proved in appendix B).

In our case, the mixing coefficients $\vec{\pi}$ are treated as parameters, and point estimations of their values are evaluated by maximizing the variational likelihood bound $\mathcal{L}\left(Q\right)$. Setting the derivative of this lower bound with respect to $\vec{\pi}$ to zero gives:

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{2.24}$$

It is noteworthy that components that provide insufficient contribution to explain the data would have their mixing coefficients driven to zero during the variational optimization. Thus, by starting with a relatively large initial value of M and then remove the redundant components after convergence, we can obtain the correct number of components. Variational learning, is able to trace the convergence systematically by monitoring the variational lower bound during the re-estimation step [33]. Indeed, at each step of the iterative re-estimation procedure, the value of this bound should never decrease. Specifically, the bound $\mathcal{L}\left(Q\right)$ is evaluated at each iteration and terminate optimization if the amount of increase from one iteration to the next is less than a threshold. For the variational inverted Dirichlet mixture model, the lower bound in (2.11) is evaluated as

$$
\begin{aligned}
\mathcal{L}(Q) &= \sum_{\mathcal{Z}} \int Q(\mathcal{Z}, \vec{\alpha}) \ln \left\{ \frac{p\left(\mathcal{X}, \mathcal{Z}, \vec{\alpha} | \vec{\pi}\right)}{Q(\mathcal{Z}, \vec{\alpha})} \right\} d\vec{\alpha} \\
&= \langle \ln p(\mathcal{X} | \mathcal{Z}, \vec{\alpha}) \rangle + \langle \ln p(\mathcal{Z} | \vec{\pi}) \rangle + \langle \ln p\left(\vec{\alpha}\right) \rangle \\
&- \langle \ln Q(\mathcal{Z}) \rangle - \langle \ln Q(\vec{\alpha}) \rangle
\end{aligned}
\tag{2.25}
$$

The variational inference for finite inverted Dirichlet mixture model can be performed via an EM-like algorithm and is summarized in Algorithm 1.

Algorithm 1 Variational learning of inverted Dirichlet mixture model

```
1: Set the initial number of components $M$.
2: Initialize the values of the hyper-parameters $u_{jl}$ and $v_{jl}$.
3: Initialize the value of $r_{ij}$ by K-means algorithm.
4: repeat
5:     The variational E-step: update the variational solutions for $Q(\mathcal{Z})$ (2.14) and $Q(\vec{\alpha})$ (2.15).
6:     The variational M-step: maximize the lower bound $\mathcal{L}(Q)$ with respect to the current value of $\vec{\pi}$ (2.24).
7: until Convergence criterion is reached.
8: Detect the optimal number of components $M$ by eliminating the components with small mixing coefficients close to 0.
```

Chapter 3

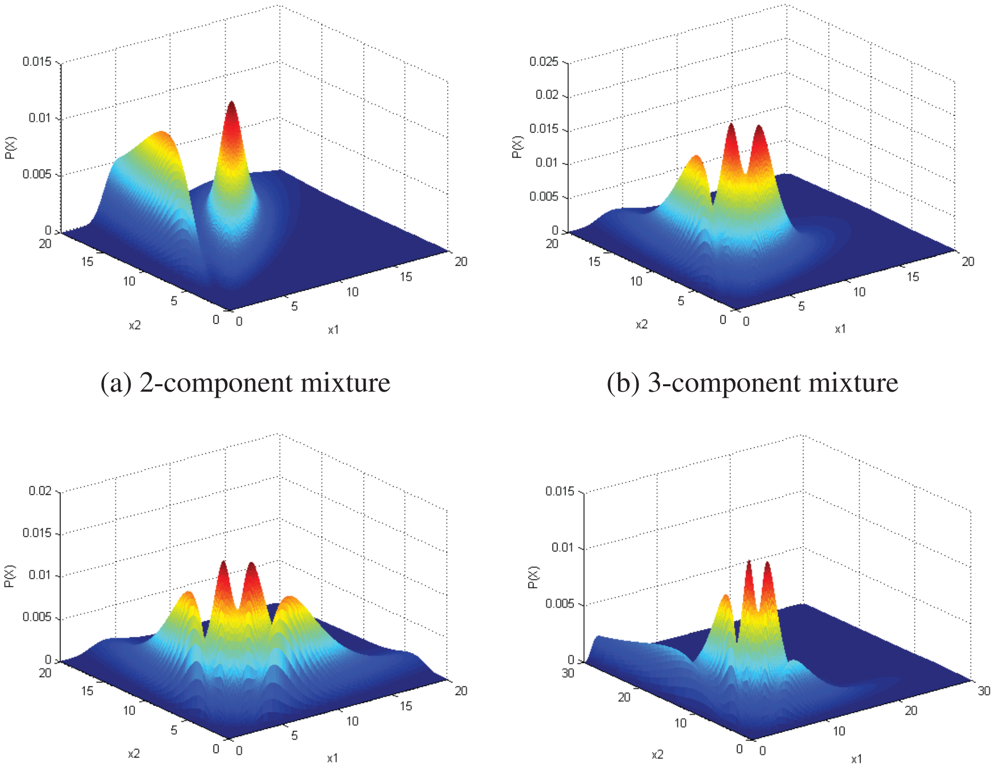
Experimental Results

3.1 Introduction

This chapter shows the experimental results of applying the proposed variational inverted Dirichlet mixture model (varIDM) on synthetic data and its applications in natural scene categorization and human activity classification. In all the experiments, the number of components M is initialized to 20 with equal mixing coefficients. The initial values of hyperparameters u_{jl} and v_{jl} were 1 and 0.01, respectively. Our experiments are performed using MATLAB on a Windows platform machine.

3.2 Synthetic Data

To show the validity of the proposed approach in parameter and model selection, it is applied on six, two-dimensional synthetic datasets. Please note that, D = 2 is chosen for ease of representation. The number of the components is set to 20 as a start point. Table 3.1 shows the real and estimated parameters, resulted from VarIDM. For estimating the number of components a threshold of $(T = 10^{-4})$ is applied to remove the redundant components that have mixing coefficients close



(a) 2-component mixture

(b) 3-component mixture

(c) 4-component mixture

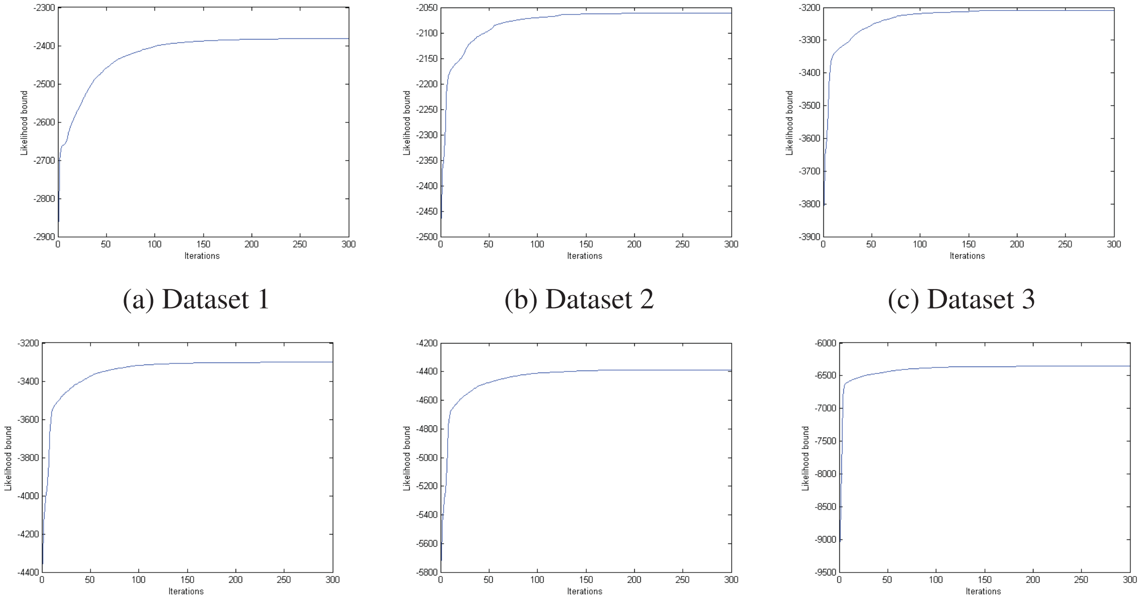
(d) 5-component mixture

Figure 3.1: Two-dimensional inverted Dirichlet mixtures

to zero. As it is shown in Table 3.1, for all the synthesized datasets, the proposed approach could successfully estimate the number of components with a good accuracy. To prove that finite inverted Dirichlet mixtures are considerably flexible and can perform in both symmetric and asymmetric modes, some examples with symmetric and asymmetric shapes are demonstrated in Figure (3.1). Figure (3.2) represents the variational lower likelihood bound. In each diagram (Figure(3.2)), the value of the likelihood bound is maximum at the point in which, the true number of components is estimated. Therefore, the variational likelihood bound can be used as a model selection criterion. In this case, there is no need to eliminate the redundant components by applying a threshold.

Table 3.1: VarIDM real and estimated parameters on different synthetic datasets. In this Table, N denotes the total number of elements, n_j shows the number of elements in cluster j. α_{j1}, α_{j2}, α_{j3}, and π_j denote the real parameters. $\bar{\alpha}_{j1}$, $\bar{\alpha}_{j2}$, $\bar{\alpha}_{j3}$, and $\bar{\pi}_j$ are the estimated parameters by variational inference.

	n_j	j	α_{j1}	α_{j2}	α_{j3}	π_j	$\bar{\alpha}_{j1}$	$\bar{\alpha}_{j2}$	$\bar{\alpha}_{j3}$	$\bar{\pi}_j$
Dataset 1 (N = 400)	200	1	20	70	4	0.50	20.47	70.70	4.07	0.5010
	200	2	40	50	5	0.50	36.42	45.16	4.62	0.4990
Dataset 2 (N = 400)	133	1	10	40	4	0.33	9.54	38.51	4.22	0.3333
	133	2	20	30	5	0.33	22.16	32.07	5.34	0.3458
	133	3	30	20	5	0.33	29.27	18.88	4.79	0.3209
Dataset 3 (N = 600)	200	1	10	40	4	0.33	9.62	38.23	3.72	0.3359
	200	2	20	30	5	0.33	18.93	28.67	4.39	0.3080
	200	3	30	20	5	0.33	29.71	19.89	5.03	0.3561
Dataset 4 (N = 600)	150	1	10	40	4	0.25	10.87	42.12	4.30	0.2446
	150	2	20	30	5	0.25	18.67	27.95	4.44	0.2710
	150	3	30	20	5	0.25	33.71	20.64	5.24	0.2533
	150	4	40	10	4	0.25	35.96	8.81	3.54	0.2311
Dataset 5 (N = 800)	200	1	10	40	4	0.25	10.43	41.66	4.01	0.2493
	200	2	20	30	5	0.25	19.14	28.42	4.81	0.2630
	200	3	30	20	5	0.25	28.42	18.66	4.49	0.2359
	200	4	40	10	4	0.25	38.04	9.30	3.86	0.2517
Dataset 6 (N = 1000)	200	1	10	40	4	0.20	8.77	39.30	3.60	0.1943
	200	2	20	30	5	0.20	17.70	29.46	4.87	0.2020
	200	3	5	60	2	0.30	4.54	56.60	2.03	0.2892
	200	4	30	20	5	0.20	28.52	19.71	4.89	0.2155
	200	5	40	10	4	0.10	36.74	9.47	3.54	0.0990



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

(f) Dataset 6

Figure 3.2: Variational lower likelihood bound in each iteration for synthesized datasets.

3.3 Natural Scene Categorization

Scene categorization is playing an important role in understanding the world through images. Human beings' brain is able to perceive complex natural scenes, understand their contents, and classify them very fast with little or no attention [34]. However, in machine vision, scene classification is a very challenging task due to the wide range of illumination, various texture and color, size and the location of the objects in the scene [2]. In this section, the proposed approach is tested on scene categorization using bag-of-visual-words representation [35, 36]. Every image contains some salient patches around the corners and the edges called keypoints, which contain valuable information about that image. Using k-means clustering these keypoints can be grouped into different clusters each of which is considered a ”visual-word”, and the group of visual-words are called visual-word vocabulary. With this definition, an image can be represented as a ”bag of visual words” [2].

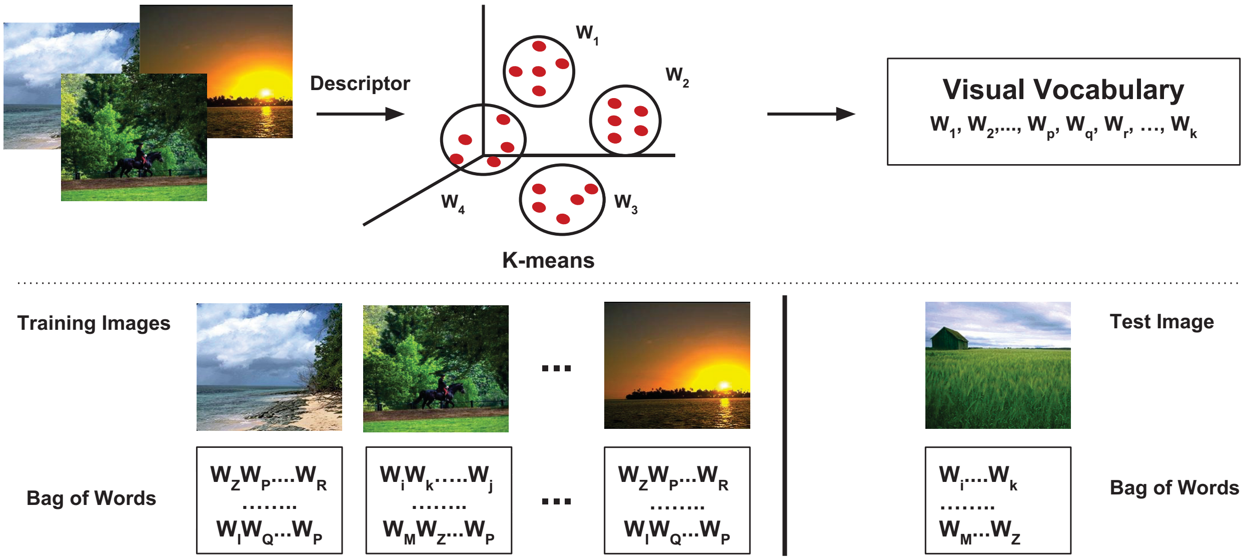

Descriptor
W_1
W_2
W_3
W_4
K-means
Visual Vocabulary
$W_1, W_2, \ldots, W_p, W_q, W_r, \ldots, W_k$
Training Images
...
Test Image
Bag of Words
$W_Z W_P \ldots W_R$ .... $W_I W_Q \ldots W_P$
$W_i W_k \ldots W_j$ .... $W_M W_Z \ldots W_P$
...
$W_Z W_P \ldots W_R$ .... $W_I W_Q \ldots W_P$
$W_i \ldots W_k$ .... $W_M \ldots W_Z$
Bag of Words

Figure 3.3: Overview of visual vocabulary formation. Figure reproduced from [2].

Before extracting the keypoints, a 5x5 window Gaussian filter with sigma = 0.5 is applied on the images to reduce the effect of noise on the extracted keypoints. The preprocessed images were fed into scale invariant feature transform (SIFT) descriptor [37] and the extracted keypoints were quantized through K-Means clustering to form our visual words. Having the visual vocabulary, each image can be represented as a d-dimensional vector containing the frequency of each visual word in that image. Figure (3.3) demonstrates this process.

For the evaluations, MIT natural scene dataset [38] is used. This dataset contains eight categories of complex scenes namely, highway (260 images), inside city (308 images), tall buildings (365 images), street (292 images), forest (328), coast (360 images), mountains (374 images), open country (410 images). The dataset is collected from COREL images and personal photographs. Based on color and texture features, the classes are divided into indoor (highway, inside city, tall buildings, street) and outdoor (forest, coast, mountains, open country) categories. Figure 3.4 shows some examples from each class.

The performance of VarIDM and GMM are visualized in confusion matrices shown in Tables (3.2) and (3.3) for indoor and outdoor categories, respectively. In each confusion matrix:

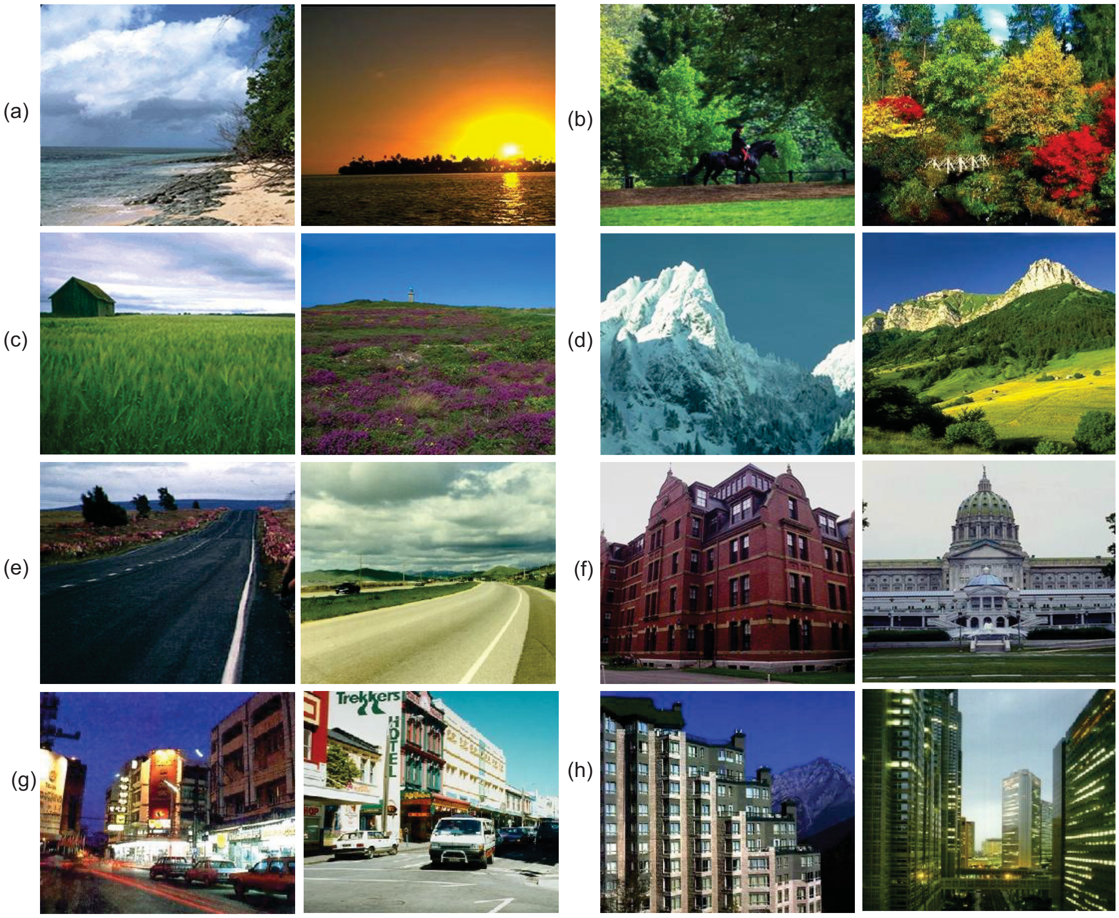


Figure 3.4: Sample frames of MIT natural scene dataset. a) Coast b) Forest c) Open country d) Mountain e) Highway f) Inside city g) Street h) Tall building.

- Each column represents the number of samples in a predicted class, and each row illustrates the samples in an actual class. In other words the value of entry $ConfMat(i, j)$ shows the number of instances which belong to class i but are classified as category j.
- The diagonal entries $ConfMat(i, j)_{i=j}$, represent the number of correctly classified samples.
- The off diagonal entries $ConfMat(i, j)_{i\neq j}$, show the system's false positives (FP) and false negatives (FN).

The same dataset was fed to Gaussian mixture model (GMM) and the results were computed.

Tables (3.4) and (3.5) illustrate the GMM's confusion matrices for indoor and outdoor datasets respectively. The overall accuracies for VarIDM and GMM are shown in Table 3.6. Running student's t-test on our results shows that VarIDM outperforms GMM at the significance level of 0.05, and p values were 0.0038 and 0.0001 for indoor and outdoor datasets respectively.

Table 3.2: VarIDM confusion matrix of indoor natural scene

	Highway	Street	Inside city	Tall building
Highway	**87**	30	6	7
Street	24	**107**	9	6
Inside city	3	7	**133**	11
Tall building	4	19	15	**145**

Table 3.3: VarIDM confusion matrix of outdoor natural scene

	Coast	Open country	Forest	Mountains
Coast	**155**	15	8	2
Open country	28	**143**	11	23
Forest	14	9	**104**	37
Mountains	3	17	32	**135**

Table 3.4: GMM confusion matrix of indoor natural scene

	Highway	Street	Inside city	Tall building
Highway	**70**	43	7	10
Street	30	**94**	14	8
Inside city	6	14	**108**	26
Tall building	5	37	16	**125**

Table 3.5: GMM confusion matrix of outdoor natural scene

	Coast	Open country	Forest	Mountains
Coast	**136**	27	11	6
Open country	39	**119**	17	30
Forest	12	19	**84**	49
Mountains	5	23	44	**115**

3.4 Human Activity Classification

Automatic human activity classification has attracted lots of attention for its important applications in surveillance for public environments such as banks and airports and subway stations, or security systems in industry and commerce for intruder detection, real-time monitoring of patients, children or elderly people, and human-computer interaction [39, 40]. Variation in environment, moving objects in the scene, camera motion, changes in scene illumination, individuals varying in posture, expression, clothing and actions, make human activity classification a challenging problem [41]. This section demonstrates our experimental results on Weizmann dataset [42] which contains 93 video sequences from nine different people, each performing ten actions namely, run, walk, skip, jumping-jack (jack), jump-forward-on-two-legs (jump), jump-in-place-on-two-legs (pjump),

Table 3.6: Average accuracy of VarIDM and GMM on indoor and outdoor natural scene dataset

	VarIDM	GMM
Outdoor	72.96	61.68
Indoor	76.99	64.76

gallop-sideways (side), wave-two-hands (wave2), wave-one-hand (wave1) and bend. The video resolution is 180×144. Samples of each class are shown in Figure (3.5). In the experiments, two different types of features are considered for video categorization. First group are local spatio-temporal features. Laptev et al. [43] detector is used to detect the space-time interest features in each video sequence. Second group are optical flow features [44], after computing the optical flow matrix on subsequent frames, a threshold (T= 0.8) considered to extract the strong optical flow responses. A mask of size 5x5 is defined around the positions with the strong optical flow values to form the total feature set. These features are fed into the minimum Redundancy Maximum Relevance (mRMR) feature selection method [45] in order to choose the most discriminative features. Using K-means algorithm, a bag of visual words is constructed, and each video is represented as a frequency histogram of the visual words. Finally, the vector of frequencies passed to VarIDM for classification. Tables (3.7) and (3.8) illustrate VarIDM and GMM confusion matrices for Weizmann dataset, respectively. We fed the same future set to Gaussian mixture model (GMM) and computed the results. The overall accuracies for VarIDM and GMM are 87.49 and 81.3 respectively. As experimental results show, the action videos can be modeled better by VarIDM rather than Gaussian mixture model. This is also illustrated by student's t-test at the significance level of 0.05 (p value=0.0005).

Table 3.7: VarIDM confusion matrix on Weizmann action dataset

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	**320**	0	0	0	0	0	0	0	0	0
Jack	0	**365**	0	0	0	0	0	0	0	0
Jump	0	0	**128**	0	25	23	53	0	0	0
Pjump	0	0	0	**269**	0	0	0	0	0	0
Run	0	0	0	0	**174**	0	22	32	0	0
Side	0	0	23	0	0	**195**	4	0	0	0
Skip	0	0	32	0	58	0	**153**	0	0	0
Walk	0	0	0	0	0	0	0	**356**	0	0
Wave1	0	0	0	0	0	0	0	0	**291**	36
Wave2	0	21	0	0	0	0	0	0	30	**261**

3.5 Comparison With Maximum Likelihood

As we mentioned in chapter one, the maximum likelihood (ML) method is the most popular approach among frequentist estimation techniques for learning a mixture model. However, it can easily get caught in saddle points or local maxima and it depends on the initially set parameters. Bdiri [1] has implemented this method via an expectation-maximization algorithm to learn the inverted Dirichlet mixture. The author tested his approach on Haberman's Survival dataset [46]. To compare the efficiency of variational learning and maximum likelihood, we applied our approach on the same dataset. Haberman's survival dataset includes 306 cases on the patients' status (survived or died) after the surgery for breast cancer. Each case has 3 features 1: Age of patient at time of surgery, 2: The year of the surgery, 3: Number of detected positive axillary nodes. In this dataset two possible status are considered for the patients:

Table 3.8: GMM confusion matrix on Weizmann action dataset

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	**320**	0	0	0	0	0	0	0	0	0
Jack	0	**331**	34	0	0	0	0	0	0	0
Jump	0	0	**116**	10	25	28	50	0	0	0
Pjump	0	0	11	**258**	0	0	0	0	0	0
Run	0	0	0	0	**160**	0	23	45	0	0
Side	0	0	47	0	0	**165**	10	0	0	0
Skip	0	0	45	0	59	0	**139**	0	0	0
Walk	0	0	0	0	11	0	0	**345**	0	0
Wave1	0	0	0	0	0	0	0	0	**259**	68
Wave2	0	23	0	0	0	0	0	0	48	**241**

- 1 : the patient survived 5 years or longer
- 2 : the patient died within 5 years

Tables 3.9, 3.10, and 3.11 show the confusion matrices of the Inverted Dirichlet mixture model and maximum likelihood (MLIDM), VarIDM and Gaussian mixture model (GMM), respectively. Both MLIDM and GMM confusion matrices are brought as reported in [1]. For the survived class, VarIDM and MLIDM results are close, but for the died class VarIDM significantly outperforms MLIDM. The overall accuracy of VarIDM, MLIDM, and GMM (Table 3.12) illustrates that VarIDM outperforms both MLIDM and GMM.

Table 3.9: MLIDM confusion matrix of Haberman's Survival dataset [1].

	Survived	Died
Survived	**195**	30
Died	51	**30**

Table 3.10: VarIDM confusion matrix of Haberman's Survival dataset.

	Survived	Died
Survived	**198**	27
Died	25	**56**

Table 3.11: GMM confusion matrix of Haberman's Survival dataset [1].

	Survived	Died
Survived	**150**	75
Died	27	**54**

Table 3.12: Overall accuracy of MLIDM, VarIDM, and GMM on Haberman's survival dataset.

Method	VarIDM	MLIDM	GMM
Accuracy	0.83	0.74	0.67

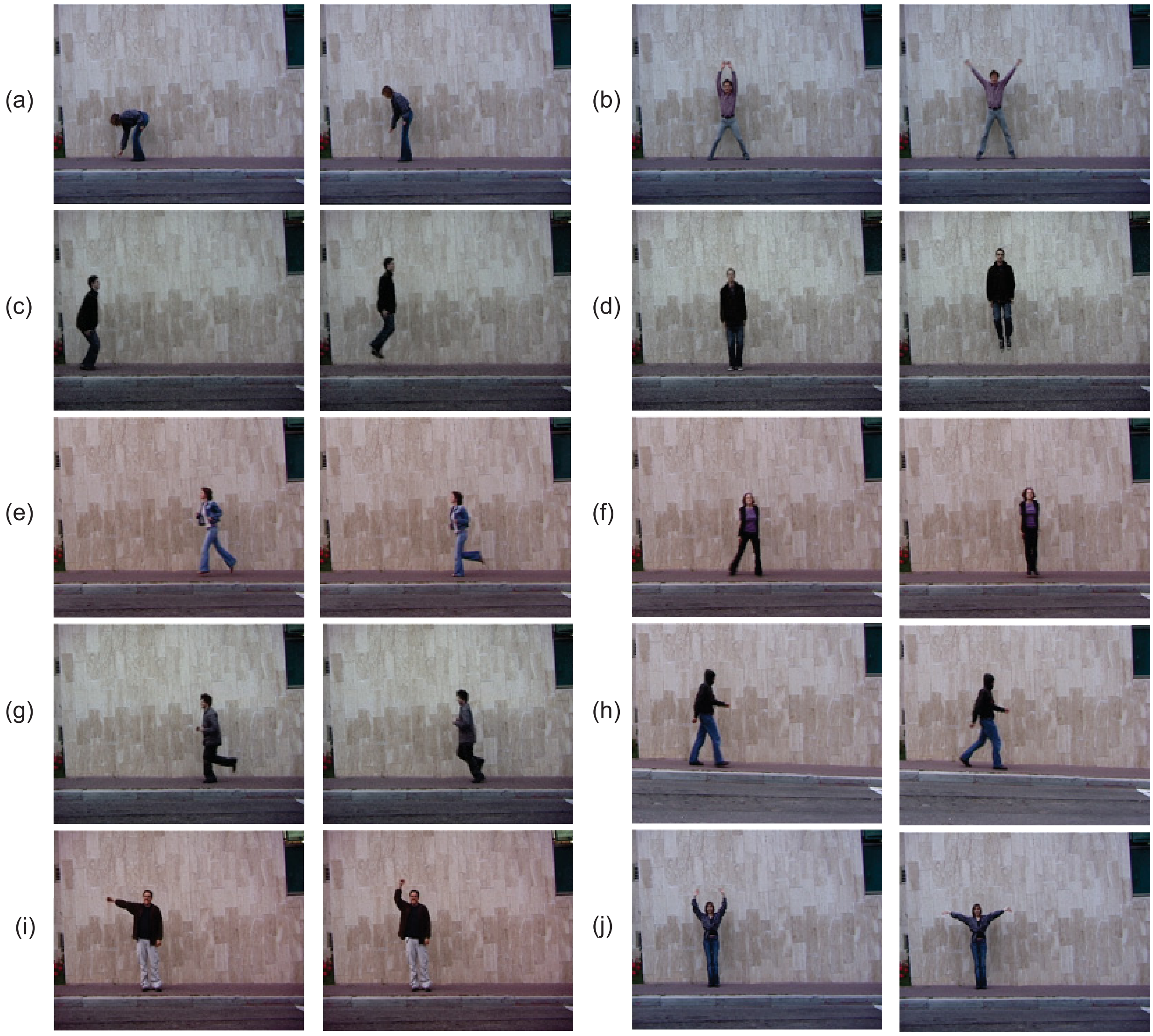


Figure 3.5: Sample frames of Weizmann dataset. a) Bend b) Jack c) Jump d) Pjump e) Run f) Side g) Skip h) Walk i) Wave1 j) Wave2.

Chapter 4

Conclusion

This thesis is mainly motivated by important role of clustering in many fields such as signal and image processing, and the related challenges in data analysis. We discussed that Gaussian mixture models are widely used in statistical modeling of the observed data. However, when it comes to modeling asymmetric data, Gaussian distribution is unable to model the data properly. For example, in image processing, the distribution of digitalized pixels is usually not symmetric [47]. The power spectrum, which is the most widely used feature in signal processing, is also asymmetric [47]. To address this problem the inverted Dirichlet mixture model is chosen due to its great ability to model both symmetric and asymmetric data. Furthermore, we defined a variational framework to learn the inverted Dirichlet mixture models. The main idea of the variational inference is approximating a complex model posterior distribution with a simpler distribution. Using this approach, both model parameters and the number of mixture components can be determined automatically and simultaneously.

The experimental results demonstrated the validity of the proposed approach in terms of parameter estimation and model selection through different synthetic datasets. Moreover, we showed the usefulness of our method on challenging real world applications, namely human activity classification which covers a wide range applications in security systems and public surveillance, and natural scene categorization that plays an important role in understanding the world through images and information retrieval. For each application, we compared the performance of our model with Gaussian mixture models in terms of accuracy and showed that the variational inverted Dirichlet model outperforms Gaussian mixture models in modeling real world data. Finally, to prove the importance of choosing a proper method to determine the number of the mixture components and to estimate the mixture parameters, we compared the variational inverted Dirichlet mixture model with IDM based on maximum likelihood approach. The test results showed that the system accuracy can improve significantly when we use variational learning.

It is worth mentioning, the variational inverted Dirichlet mixture models can be used in many other applications that are dealing with asymmetric data. There are several promising avenues for future research. For instance, the covariance structure imposed by the inverted Dirichlet is strictly positive. This is rather restrictive and does not allow for modeling the data in a flexible way. Thus, it is possible to consider the generalized inverted Dirichlet to enlarge the applicability of the proposed model.

Appendix A

Proof of Equations (2.14) and (2.15)

The general expression for the variational solution $Q_s(\Theta_s)$ (Eq. (2.13)), can be written as:

$$\ln Q_s(\Theta_s) = \langle \ln p(\mathcal{X}, \Theta) \rangle_{j \neq s} + const \tag{A.1}$$

where *const* is an additive constant denoting all the terms that are independent of $Q_s(\Theta_s)$. Considering the joint distribution represented in Eq. (2.9), the following variational solutions for $Q(\mathcal{Z})$ and $Q(\vec{\alpha})$ can be developed.

A.1 Proof of Eq. (2.14): Variational Solution to $Q(\mathcal{Z})$

$$\ln Q(Z_{ij}) = Z_{ij} \left[\ln \pi_j + \mathcal{R}_j + \sum_{l=1}^{D+1} (\bar{\alpha}_{jl} - 1) \ln X_{il} \right] + const \tag{A.2}$$

where

$$\mathcal{R}_j = \left\langle \ln \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} \right\rangle_{\alpha_{j1}, \ldots, \alpha_{jD+1}} \tag{A.3}$$

and

$$\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}}{v_{jl}} \tag{A.4}$$

Since a closed-form expression cannot be found for $\mathcal{R}_j$, the standard variational inference cannot be applied directly. Therefore, to obtain a closed-form expression, a lower bound approximation is proposed. To provide traceable approximations, the second-order Taylor series expansion is applied in variational inference [47, 48]. In fact, the function $\mathcal{R}_j$ is approximated using a second-order Taylor expansion about the expected values of the parameters $\vec{\alpha}_j$. Here, $\widetilde{\mathcal{R}_j}$ is defined to denote the approximation of $\mathcal{R}_j$, and $(\bar{\alpha}_{j1}, ..., \bar{\alpha}_{jD+1})$ to represent the expected values of $\vec{\alpha}_j$. Replacing $\mathcal{R}_j$ by $\widetilde{\mathcal{R}}_j$ makes the optimization in Eq.(A.2) traceable.

Considering the logarithmic form of Eq. (2.6) , formula A.2 can be written as

$$\ln Q(\mathcal{Z}) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \ln \rho_{ij} + const \tag{A.5}$$

where

$$\ln \rho_{ij} = \ln \pi_j + \widetilde{\mathcal{R}}_j + \sum_{l=1}^{D} (\bar{\alpha}_{jl} - 1) \ln X_{il} \tag{A.6}$$

All the terms, independent of Z_{ij} can be added to the constant part, therefore, it is straight forward to show

$$Q(\mathcal{Z}) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \rho_{ij}^{Z_{ij}} \tag{A.7}$$

In order to find the exact formula for $Q(\mathcal{Z})$ the Eq. (A.7) needs to be normalized. Simple calculations lead to

$$Q(\mathcal{Z}) = \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{A.8}$$

where

$$r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M} \rho_{ij}} \tag{A.9}$$

Note that $\sum_{j=1}^{M} r_{ij} = 1$, therefore, the standard result for $Q(\mathcal{Z})$ will be

$$\langle Z_{ij} \rangle = r_{ij} \tag{A.10}$$

A.2 Proof of Eq. (2.15): Variational Solution to $Q(\vec{\alpha})$

Having a mixture model with M components and assuming the parameter α_{jl} are independent, $Q(\vec{\alpha})$ can be factorized as

$$Q(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D+1} Q(\alpha_{jl}) \tag{A.11}$$

Consider the variational optimization for the specific factor $Q(\alpha_{js})$. The logarithm of the optimized factor is given by

$$\ln Q(\alpha_{js}) = \sum_{i=1}^{N} r_{ij} \mathscr{T}(\alpha_{js}) + \alpha_{js} \sum_{i=1}^{N} r_{ij} \ln X_{is} + (u_{js} - 1) \ln \alpha_{js} - v_{js} \alpha_{js} + const \tag{A.12}$$

where

$$\mathscr{T}(\alpha_{js}) = \left\langle \ln \frac{\Gamma\big(\alpha_s + \sum_{l \neq s}^{D+1} \alpha_{jl}\big)}{\Gamma(\alpha_s) \prod_{l \neq s}^{D+1} \Gamma(\alpha_{jl})} \right\rangle_{\Theta \neq \alpha_{js}} \tag{A.13}$$

$\mathscr{T}$ is a function of α_{js}. Since $\mathscr{T}(\alpha_{js})$ is intractable, a lower bound estimation should be found for it. Hence, a first-order Taylor expansion [47] around $\bar{\alpha}_{js}$ (the expected value of α_{js}) is used (See Appendix B).

$$\mathscr{T}(\alpha_{js}) \geq \bar{\alpha}_{js} \ln \alpha_{js} \left\{ \Psi\Big(\sum_{l=1}^{D+1} \bar{\alpha}_{jl} \Big) - \Psi(\bar{\alpha}_{js}) + \sum_{l \neq s}^{D+1} \bar{\alpha}_{jl} \times \Psi'\Big(\sum_{l=1}^{D+1} \bar{\alpha}_{jl} \Big) (\langle \ln \alpha_{jl} \rangle - \ln \bar{\alpha}_{jl}) \right\} + const \tag{A.14}$$

Substituting this lower bound into Eq. (A.12), results in an optimal solution for α_{js}

$$
\begin{aligned}
\ln Q(\alpha_{js}) = & \sum_{i=1}^{N} r_{ij}\bar{\alpha}_{js}\ln\alpha_{js}\Big[\Psi\big(\sum_{i=1}^{D+1}\bar{\alpha}_{jl}\big) - \Psi(\bar{\alpha}_{js}) + \sum_{l\neq s}^{D+1}\Psi'\big(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}\big)\bar{\alpha}_{jl}(\langle\ln\alpha_{jl}\rangle - \ln\bar{\alpha}_{jl})\Big] \\
& + \alpha_{js}\sum_{i=1}^{N} r_{ij}\ln X_{is} + (u_{js}-1)\ln\alpha_{js} - v_{js}\alpha_{js} + const \\
& = \ln\alpha_{js}(u_{js}+\phi_{js}-1) - \alpha_{js}(v_{js}-\nu_{js}) + const
\end{aligned} \tag{A.15}
$$

where

$$
\phi_{js} = \sum_{i=1}^{N} r_{ij}\bar{\alpha}_{js}\Bigg[\Psi\big(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}\big) - \Psi(\bar{\alpha}_{js}) + \sum_{l\neq s}^{D+1}\Psi'\big(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}\big) \times \bar{\alpha}_{jl}(\langle\ln\alpha_{jl}\rangle - \ln\bar{\alpha}_{jl})\Bigg] \tag{A.16}
$$

$$
\nu_{js} = \sum_{i=1}^{N} r_{ij}\ln X_{is} \tag{A.17}
$$

By taking the exponential of Eq. (A.15) which is the logarithmic form of Gamma distribution, we will have

$$
Q(\alpha_{js}) \propto \alpha_{js}^{u_{js}+\phi_{js}-1} e^{-(v_{js}-\nu_{js})\alpha_{js}} \tag{A.18}
$$

The optimal solution to the parameters are as

$$
\begin{aligned}
u_{js}^{*} &= u_{js} + \phi_{js} \\
v_{js}^{*} &= v_{js} - \nu_{js}
\end{aligned} \tag{A.19}
$$

Appendix B

Proof of Equations (2.18) and (A.14)

B.1 Lower bound of $\mathcal{R}_j$: Proof of Eq.(2.18)

Since $\mathcal{R}_j$ in Eq.(A.3) is intractable, a non-linear approximation of the lower bound of $\mathcal{R}_j$ is calculated using the second-order Taylor expansion. First, function $\mathcal{H}$ is defined as follow

$$\mathcal{H}(\vec{\alpha}_j) = \mathcal{H}(\alpha_{j1}, \dots, \alpha_{jD+1}) = \ln \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} \tag{B.1}$$

where $\alpha_{jl} > 1$. Using the second-order Taylor expansion for $\ln \vec{\alpha}_j = (\ln \alpha_{j1}, \dots, \ln \alpha_{jD+1})$ around $\ln \vec{\alpha}_{j,0} = (\ln \vec{\alpha}_{j1,0}, \dots, \ln \vec{\alpha}_{jD+1,0})$, the lower bound of $\mathcal{H}(\vec{\alpha}_j)$ is obtained as

$$\mathcal{H}(\vec{\alpha}_j) \geq \mathcal{H}(\vec{\alpha}_{j,0}) + (\ln \vec{\alpha}_j - \ln \vec{\alpha}_{j,0})^T \nabla \mathcal{H}(\vec{\alpha}_{j,0}) + \frac{1}{2!}(\ln \vec{\alpha}_j - \ln \vec{\alpha}_{j,0})^T \nabla^2 \mathcal{H}(\vec{\alpha}_{j,0})(\ln \vec{\alpha}_j - \ln \vec{\alpha}_{j,0}) \tag{B.2}$$

where $\nabla \mathcal{H}(\vec{\alpha}_{j,0})$ is the gradient of $\mathcal{H}$ at $\vec{\alpha}_j = \vec{\alpha}_{j,0}$ and $\nabla^2 \mathcal{H}(\vec{\alpha}_{j,0})$ is the Hessian matrix, which gives

$$\begin{aligned}\mathcal{H}(\vec{\alpha}_j) \geq \mathcal{H}(\vec{\alpha}_{j,0}) &+ \sum_{l=1}^{D+1} \frac{\partial \mathcal{H}(\vec{\alpha}_j)}{\partial \ln \alpha_{jl}} |_{\vec{\alpha}_j = \vec{\alpha}_{j,0}} (\ln \alpha_{jl} - \ln \alpha_{jl,0}) \\ &+ \frac{1}{2} \sum_{a=1}^{D+1} \sum_{b=1}^{D+1} \frac{\partial^2 \mathcal{H}(\vec{\alpha}_{j,0})}{\partial \ln \alpha_{ja} \partial \ln \alpha_{jb}} |_{\vec{\alpha}_j = \vec{\alpha}_{j,0}} (\ln \alpha_{ja} - \ln \alpha_{ja,0}) \\ &\times (\ln \alpha_{jb} - \ln \alpha_{jb,0})\end{aligned} \tag{B.3}$$

Taking the expectation of Eq.(B.3), the lower bound of function $\mathcal{R}_j$ will be

$$\begin{aligned}\mathcal{R}_j \geq \widetilde{\mathcal{R}}_j = \ln &\frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl,0})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl,0})} \\ &+ \sum_{l=1}^{D+1} \alpha_{jl,0} \left[\Psi\left(\sum_{l=1}^{D+1} \alpha_{jl,0} \right) - \Psi(\alpha_{jl,0}) \right] \times [\langle \ln \alpha_{jl} \rangle - \ln \alpha_{jl,0}] \\ &+ \frac{1}{2} \sum_{l=1}^{D+1} \alpha_{jl,0}^2 \left[\Psi'\left(\sum_{l=1}^{D+1} \alpha_{jl,0} \right) - \Psi'(\alpha_{jl,0}) \right] \times \langle (\ln \alpha_{jl} - \ln \alpha_{jl,0})^2 \rangle \\ &+ \frac{1}{2} \sum_{a=1}^{D+1} \sum_{\substack{b=1 \\ (b \neq a)}}^{D+1} \left\{ \alpha_{ja,0} \alpha_{jb,0} \Psi'\left(\sum_{l=1}^{D+1} \alpha_{jl,0} \right) (\langle \ln \alpha_{ja} \rangle - \ln \alpha_{ja,0}) \right. \\ &\left. \times (\langle \ln \alpha_{jb} \rangle - \ln \alpha_{jb,0}) \right\}\end{aligned} \tag{B.4}$$

To prove that the second-order Taylor expansion of $\mathcal{H}(\vec{\alpha}_j)$ is a lower bound of $\mathcal{H}(\vec{\alpha}_j)$, it is shown that $\Delta\mathcal{H}(\vec{\alpha}_j) \geq 0$, where $\Delta\mathcal{H}(\vec{\alpha}_j)$ denotes the difference between $\mathcal{H}(\vec{\alpha}_j)$ and its second-order Taylor expansion. The Hessian of $\Delta\mathcal{H}(\vec{\alpha}_j)$ with respect to $(\ln \alpha_{jl}, \ldots, \ln \alpha_{jD+1})$ is given by (B.5).

$$Hess = \begin{pmatrix} \begin{array}{l} \alpha_{j1}[\Psi(\sum_{l=1}^{D+1}\alpha_{jl}) - \Psi(\alpha_{j1})] \\ +\alpha_{j1}^2[\Psi'(\sum_{l=1}^{D+1}\alpha_{jl}) - \Psi'(\alpha_{jl})] \\ -\bar{\alpha}_{j1}^2[\Psi'(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}) - \Psi'(\bar{\alpha}_{j1})] \end{array} & \cdots & \begin{array}{r} \alpha_{j1}\alpha_{jD+1}\Psi'(\sum_{l=1}^{D+1}\alpha_{jl}) \\ -\bar{\alpha}_{j1}\bar{\alpha}_{jD+1}\Psi'(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}) \end{array} \\ \vdots & \ddots & \vdots \\ \begin{array}{r} \alpha_{j1}\alpha_{jD+1}\Psi'(\sum_{l=1}^{D+1}\alpha_{jl}) \\ -\bar{\alpha}_{j1}\bar{\alpha}_{jD+1}\Psi'(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}) \end{array} & \cdots & \begin{array}{l} \alpha_{jD+1}[\Psi(\sum_{l=1}^{D+1}\alpha_{jl}) - \Psi(\alpha_{jD+1})] \\ +\alpha_{jD+1}^2[\Psi'(\sum_{l=1}^{D+1}\alpha_{jl}) - \Psi'(\alpha_{jD+1})] \\ -\bar{\alpha}_{jD+1}^2[\Psi'(\sum_{l=1}^{D+1}\bar{\alpha}_{jl}) - \Psi'(\bar{\alpha}_{jD+1})] \end{array} \end{pmatrix} \tag{B.5}$$

Substituting $(\ln\alpha_{j1}, \dots, \ln\alpha_{jD+1})$ by the critical point $(\ln\alpha_{j1,0}, \dots, \ln\alpha_{jD+1,0})$, reduces (B.5) to a positive-definite diagonal matrix. Since $(\ln\alpha_{j1,0}, \dots, \ln\alpha_{jD+1,0})$ is the only critical point, and for all $\alpha_{jl} > 1$, $\Delta\mathcal{H}(\vec{\alpha}_j)$ is continuous and differentiable, the critical point $(\ln\alpha_{j1,0}, \dots, \ln\alpha_{jD+1,0})$ is also the global minimum of $\Delta\mathcal{H}(\vec{\alpha}_j)$. When $(\ln\alpha_{j1}, \dots, \ln\alpha_{jD+1}) = (\ln\alpha_{j1,0}, \dots, \ln\alpha_{jD+1,0})$, the global minimum value 0, is reached, therefore, the second-order Taylor expansion is definitely a lower bound of $\mathcal{H}$.

B.2 Lower Bound of $\mathscr{T}(\alpha_{js})$: Proof of(A.14)

The lower bound of $\mathscr{T}(\alpha_{js})$ is approximated in [49] by a first-order Taylor expansion. The first-order Taylor expansion of a convex function is a tangent line of that function at a specific value. Function $\mathcal{F}(\alpha_{js})$ is defined as

$$\mathcal{F}(\alpha_{js}) = \ln \frac{\Gamma\Big(\alpha_{js} + \sum_{l\neq s}^{D+1}\alpha_{jl}\Big)}{\Gamma(\alpha_{js})\prod_{l\neq s}^{D+1}\Gamma(\alpha_{jl})} \tag{B.6}$$

B.2.1 Convexity of $\mathcal{F}(\alpha_{js})$

Since it cannot directly be shown that $\mathcal{F}(\alpha_{js})$ is a convex function of α_{js}, a relative convexity similar to [47] is considered. It can be demonstrated that $\mathcal{F}(\alpha_{js})$ is convex relative to $\ln \alpha_{js}$. Function $\mathcal{F}$, is considered to be convex on an interval if and only if its second derivative is nonnegative in that interval. The first and second derivatives of $\mathcal{F}(\alpha_{js})$ with respect to $\ln \alpha_{js}$ are

$$\frac{\partial \mathcal{F}(\alpha_{js})}{\partial \ln \alpha_{js}} = \left[\Psi\bigg(\alpha_{js} + \sum_{l \neq s}^{D+1} \alpha_{jl}\bigg) - \Psi(\alpha_{js})\bigg)\right]\alpha_{js} \tag{B.7}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{F}(\alpha_{js})}{\partial (\ln \alpha_{js})^2} &= \left[\Psi\bigg(\alpha_{js} + \sum_{l \neq s}^{D+1} \alpha_{jl}\bigg) - \Psi(\alpha_{js})\bigg)\right]\alpha_{js} + \left[\Psi'\bigg(\alpha_{js} + \sum_{l \neq s}^{D+1} \alpha_{jl}\bigg) - \Psi'(\alpha_{js})\bigg)\right]\alpha_{js}^2 \\ &= \alpha_{js} \int_0^\infty \frac{1 - e^{-(\sum_{l \neq s}^{D+1} \alpha_{jl})t}}{1 - e^{-t}} e^{-\alpha_{js} t}(1 - \alpha_{js} t) dt\end{aligned} \tag{B.8}$$

The integral representation of $\Psi(x)$ and $\Psi'(x)$ are defined by

$$\Psi(x) = \int_0^\infty \left(\frac{e^{-t}}{t} - \frac{e^{-xt}}{1 - e^{-t}}\right) dt \tag{B.9}$$

$$\Psi'(x) = \int_0^\infty \frac{t e^{-xt}}{1 - e^{-t}} dt \tag{B.10}$$

Considering (B.9) and (B.10), Eq.(B.8) can be written as

$$\frac{\partial^2 \mathcal{F}(\alpha_{js})}{\partial (\ln \alpha_{js})^2} = \alpha_{js} \int_0^\infty f_1(t) f_2(t) dt \tag{B.11}$$

where $f_1(t)$ and $f_2(t)$ are

$$f_1(t) = \frac{1 - e^{-(\sum_{l \neq s}^{D+1} \alpha_{jl})t}}{1 - e^{-t}} \tag{B.12}$$

$$f_2(t) = e^{-\alpha_{js} t}(1 - \alpha_{js} t) \tag{B.13}$$

when $\sum_{l\neq s}^{D+1} \alpha_{jl} > 1$

- if $t > \frac{1}{\alpha_{js}}$ then $f_1(t) < f_1(\frac{1}{\alpha_{js}})$, and $f_2(t) < 0$
- if $t < \frac{1}{\alpha_{js}}$ then $f_1(t) > f_1(\frac{1}{\alpha_{js}})$, $f_2(t) > 0$

Therefore Eq.(B.11) can be rewritten as

$$\begin{aligned}
\frac{\partial^2 \mathcal{F}(\alpha_{js})}{\partial (\ln \alpha_{js})^2} &= \alpha_{js} \left\{ \int_0^{\frac{1}{\alpha_{js}}} f_1(t) f_2(t) dt + \int_{\frac{1}{\alpha_{js}}}^{\infty} f_1(t) f_2(t) dt \right\} \\
&> \alpha_{js} \left\{ \int_0^{\frac{1}{\alpha_{js}}} f_1(\frac{1}{\alpha_{js}}) f_2(t) dt + \int_{\frac{1}{\alpha_{js}}}^{\infty} f_1\left(\frac{1}{\alpha_{js}}\right) f_2(t) dt \right\} \\
&= \alpha_{js} f_1\left(\frac{1}{\alpha_{js}}\right) \int_0^{\infty} f_2(t) dt \\
&= \alpha_{js} f_1\left(\frac{1}{\alpha_{js}}\right) \lim_{t\to\infty} t e^{-\alpha_{js} t} = 0
\end{aligned} \tag{B.14}$$

Hence, it is proven, when $\sum_{l\neq s}^{D+1} \alpha_{jl} > 1$, $\mathcal{F}(\alpha_{js})$ is convex relative to $\ln \alpha_{js}$.

B.2.2 Evaluating Lower Bound by First-order Taylor Expansion

The lower bound of $\mathcal{F}(\alpha_{js})$ can be calculated by applying the first-order Taylor expansion of $\mathcal{F}(\alpha_{js})$ for $\ln \alpha_{js}$ at $\ln \alpha_{js,0}$, since is a convex function relative to $\ln \alpha_{js}$.

$$\begin{aligned}\mathcal{F}(\alpha_{js}) &\geq \mathcal{F}(\alpha_{js,0}) + \frac{\partial \mathcal{F}(\alpha_{js})}{\partial \ln \alpha_{js}}|\alpha_{js} = \alpha_{js,0}(\ln \alpha_{js} - \ln \alpha_{js,0}) \\ &= \mathcal{F}(\alpha_{js,0}) + \frac{\partial \mathcal{F}(\alpha_{js})}{\partial \alpha_{js}} \frac{\partial \alpha_{js}}{\partial \ln \alpha_{js}}|\alpha_{js} = \alpha_{js,0}(\ln \alpha_{js} - \ln \alpha_{js,0}) \\ &= \ln \frac{\Gamma\Big(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl}\Big)}{\Gamma(\alpha_{js,0}) \prod_{l\neq s}^{D+1} \Gamma(\alpha_{jl})} + \left[\Psi\Big(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl}\Big) - \Psi(\alpha_{js,0})\Big)\right]\alpha_{js,0}(\ln \alpha_{js} - \ln \alpha_{js,0})\end{aligned} \tag{B.15}$$

Note that when $\alpha_{js} = \bar{\alpha}_{js}$, the equality is reached. Substituting (B.15) in (A.14) will result in

$$\begin{aligned}\mathscr{T}(\alpha)_{js} \geq &\left\langle \ln \frac{\Gamma\Big(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl}\Big)}{\Gamma(\alpha_{js,0}) \prod_{l\neq s}^{D+1} \Gamma(\alpha_{jl})} + \left[\Psi\Big(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl}\Big) - \Psi(\alpha_{js,0})\right]\alpha_{js,0}(\ln \alpha_{js} - \ln \alpha_{js,0}) \right\rangle_{\vec{\alpha}\neq\alpha_{js}} \\ &\ln \alpha_{js}\alpha_{js,0}\left\{\left\langle \Psi\Big(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl}\Big)\right\rangle_{\vec{\alpha}\neq\alpha_{js}} - \Psi(\alpha_{js,0})\right\} + const\end{aligned} \tag{B.16}$$

The calculation of the expectation $\langle \Psi(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl})\rangle_{\vec{\alpha}\neq\alpha_{js}}$ (in (B.16)) is analytically intractable. With the same approach explained in section (B.2.1), it can be inferred, for $l = \{1, \ldots, D+1\}$ and $l \neq s$, $\Psi(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl})$ is a convex function relative to $\ln \alpha_{js,0}$. To calculate the lower bound, a first-order Taylor expansion for the function $\Psi(\sum_{i=1}^{n} x_i + y)$ at $\ln \hat{x}, \hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ is applied

$$\Psi\Big(\sum_{i=1}^{n} x_i + y\Big) \geq \Psi\Big(\sum_{i=1}^{n} \hat{x}_i + y) + \sum_{i=1}^{n} (\ln x_i - \ln \hat{x}_i)\Psi'(\sum_{i=1}^{n} \hat{x}_i + y)\hat{x}_i \tag{B.17}$$

Considering (B.17), the approximation lower bound of $\langle \Psi(\alpha_{js,0} + \sum_{l\neq s}^{D+1} \alpha_{jl})\rangle_{\vec{\alpha}\neq\alpha_{js}}$ is given by

$$\left\langle \Psi\left(\sum_{l\neq s}^{D+1} \alpha_{jl} + \alpha_{js,0}\right)\right\rangle_{\vec{\alpha}\neq\alpha_{js}} \geq \Psi\left(\sum_{l=1}^{D+1} \alpha_{jl,0}\right) + \sum_{l\neq s}^{D+1} \alpha_{jl,0}\Psi'\left(\sum_{l=1}^{D+1} \alpha_{jl,0}\right)(\langle \ln \alpha_{jl}\rangle - \ln \alpha_{jl,0}) \tag{B.18}$$

The lower bound of $\mathscr{T}(\alpha_{js})$ can be calculated by substituting (B.18) to (A.14)

$$\mathscr{T}(\alpha_{js}) \geq \ln \alpha_{js}\alpha_{js,0}\left\{\Psi\left(\sum_{l=1}^{D+1} \alpha_{jl,0}\right) - \Psi(\alpha_{js,0}) + \sum_{l\neq s}^{D+1} \alpha_{jl,0}\Psi'\left(\sum_{l=1}^{D+1} \alpha_{jl,0}\right)(\langle \ln \alpha_{jl}\rangle - \ln \alpha_{jl,0})\right\} + const \tag{B.19}$$

List of References

[1] T. Bdiri and N. Bouguila. Positive vectors clustering using inverted dirichlet finite mixture models. *Expert Syst. Appl.*, 39:1869–1882, 2012.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV'06, pages 517–530, Berlin, Heidelberg, 2006. Springer-Verlag.

[3] R.W. Picard. Light-years from lena: video and image libraries of the future. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 310–313, 1995.

[4] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang. Supporting ranked boolean similarity queries in mars. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):905–925, 1998.

[5] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26:35 – 58, 2001.

[6] N.R. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30:847 – 857, 1997.

[7] Comaniciu, D. and Meer P. Distribution Free Decomposition of Multivariate Data. *Pattern Analysis & Applications*, 2:22–30, 1999.

[8] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *Proc. of the Nineteenth International Conference on Machine Learning (ICML)*, pages 27–34, 2002.

[9] Dougherty, E. R. and Brun, M. A Probabilistic Theory of Clustering. *Pattern Recognition*, 37:917–925, 2004.

[10] A. M. Bagirov, J. Ugon, and D. Webb. Fast modified global k-means algorithm for incremental cluster construction. *Pattern Recognition*, 44:866–876, 2011.

[11] M.H.C. Law, A.P. Topchy, and A.K. Jain. Multiobjective data clustering. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 424–430, 2004.

[12] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58:155–176, 1996.

[13] V. Garcia, F. Nielsen, and R. Nock. Levels of details for gaussian mixture models. In *ACCV (2)*, pages 514–525, 2009.

[14] M. Dixit, N. Rasiwasia, and N. Vasconcelos. Adapted gaussian models for image classification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 937–943, 2011.

[15] A. M. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *Computer Vision - ECCV 2000, 6th European Conference on Computer Vision, Dublin, Ireland, June 26 - July 1, 2000, Proceedings, Part II*, pages 751–767, 2000.

[16] L. Liu and G. Fan. Combined key-frame extraction and object-based video segmentation. *IEEE Trans. Circuits Syst. Video Techn.*, 15:869–884, 2005.

[17] X. Song and G. Fan. Joint key-frame extraction and object segmentation for content-based video analysis. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16:904–914, 2006.

[18] M. S. Allili, N. Bouguila, and D. Ziou. Finite generalized gaussian mixture modeling and applications to image and video foreground segmentation. In *Proc. of the Fourth Canadian Conference on Computer and Robot Vision (CRV)*, pages 183–190, 2007.

[19] N. Bouguila. Spatial color image databases summarization. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, pages 953–956, 2007.

[20] N. Bouguila and D. Ziou. Online clustering via finite mixtures of dirichlet and minimum message length. *Eng. Applications of Artificial Intelligence*, 19(4):371–379, 2006.

[21] T. Bdiri and N. Bouguila. Bayesian learning of inverted dirichlet mixtures for SVM kernels generation. *Neural Computing and Applications*, 23:1443–1458, 2013.

[22] T. Bdiri and N. Bouguila. Learning inverted dirichlet mixtures for positive data clustering. In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - 13th International Conference, RSFDGrC 2011, Moscow, Russia, June 25-27, 2011. Proceedings*, pages 265–272, 2011.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[24] Z. Ghahramani and M. J. Beal. Variational inference for bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems (NIPS)*, pages 449–455, 1999.

[25] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[26] M. Opper and G. Sanguinetti. Variational inference for markov jump processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[27] G. G. Tiao and I. Cuttman. The inverted dirichlet distribution with applications. *Journal of the American Statistical Association*, 60:793–805, 1965.

[28] H. Yassaee. Inverted dirichlet distribution and multivariate logistic distribution. *Canadian Journal of Statistics*, 2:99–105, 1974.

[29] M. Ghorbel. On the inverted dirichlet distribution. *Communications in Statistics - Theory and Methods*, 39:21–37, 2009.

[30] A. Corduneanu and Ch. M. Bishop. Variational Bayesian model selection for mixture distributions. In *Proceedings of Artificial Intelligence and Statistics 2001*, pages pp. 27–34, 2001.

[31] L. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems 8*, pages 486–492. MIT Press, 1995.

[32] T. S. Jaakkola and M. I. Jordan. Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, pages 340–348, 1996.

[33] H. Attias. A variational bayesian framework for graphical models. In *In Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[34] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531, 2005.

[35] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on*

Workshop on Multimedia Information Retrieval, pages 197–206, New York, NY, USA, 2007. ACM.

[36] M. E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1447–1454, 2006.

[37] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[38] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, 2001.

[39] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43:16:1–16:43, April 2011.

[40] J. Luo, W. Wang, and H. Qi. Feature extraction and representation for distributed multi-view human action recognition. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 3:145–154, 2013.

[41] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra. Effective codebooks for human action representation and classification in unconstrained videos. *IEEE Transactions on Multimedia*, 14:1234–1245, 2012.

[42] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29:2247–2253, 2007.

[43] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[44] B. K.P. Horn and B. G. Schunck. Determining optical flow. Technical report, Cambridge, MA, USA, 1980.

[45] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1226–1238, 2005.

[46] T. S. Lim. UCI machine learning repository, 1999.

[47] Z. Ma and A. Leijon. Bayesian estimation of beta mixture models with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33:2160–2173, 2011.

[48] M. W. Woolrich and Behrens T. E. Variational bayes inference of spatial mixture models for segmentation. *Medical Imaging, IEEE Transactions on*, 25:1380–1391, 2006.

[49] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37:183–233, 1999.