EFFICIENT COMPUTATION OF LOG-LIKELIHOOD FUNCTION IN CLUSTERING OVERDISPERSED COUNT DATA

Masoud Daghyani

A thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science
(Electrical and Computer Engineering)
Concordia University
Montréal, Québec, Canada

September 2019

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Masoud Daghyani**

Entitled: **Efficient Computation of Log-likelihood Function in Clustering Overdispersed Count Data**

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

Master of Applied Science
(Electrical and Computer Engineering)

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

Signed by the final examining commitee:

Dr. Hassan Rivaz (ECE) ____________________ Chair

Prof. Nizar Bouguila (CIISE) ____________________ Supervisor

Dr. Hassan Rivaz (ECE) ____________________ Internal Examiner

Dr. Farjad Shadmehri (MIE) ____________________ External Examiner

Dr. Fereshteh Mafakheri (CIISE) ____________________ External Examiner

Approved by ____
Prof. Yousef R. Shayan, Chair, Electrical and Computer Engineering

2019.08.14 ____

Prof. Amir Asif, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Efficient Computation of Log-likelihood Function in Clustering Overdispersed Count Data

Masoud Daghyani

In this work, we present an overdispersed count data clustering algorithm, which uses the mesh method for computing the log-likelihood function, of the multinomial Dirichlet, multinomial generalized Dirichlet, and multinomial Beta-Liouville distributions. Count data are often used in many areas such as information retrieval, data mining, and computer vision. The multinomial Dirichlet distribution (MDD) is one of the widely used methods of modeling multi-categorical count data with overdispersion. In recent works, the use of the mesh algorithm, which involves the approximation of the multinomial Dirichlet distribution's (MDD) log-likelihood function, based on the Bernoulli polynomials; has been proposed instead of using the traditional numerical computation of the log-likelihood function which either results in instability, or leads to long run times that make its use infeasible when modeling large-scale data. Therefore, we extend the mesh algorithm approach for computing the log likelihood function of more flexible distributions, namely multinomial generalized Dirichlet (MGD) and multinomial Beta-Liouville (MBL). A finite mixture model based on these distributions, is optimized by expectation maximization, and attempts to achieve a high accuracy for count data clustering. Through a set of experiments, the proposed approach shows its merits in two real-world clustering problems, that concern natural scenes categorization and facial expression recognition.

Acknowledgments

I would like to express my deepest gratitude to my supervisor Prof. Nizar Bouguila. It is very difficult to put into words to exactly express my appreciation to him as an distinguished scientist with a wonderful personality. It is one of my greatest honours in my entire life to be his student. As an extraordinary supervisor, he trusted me and allowed me to start my research under his excellent supervision. The completion of this thesis was not truly possible without his kind, patience, wise guidance, endless supports, and never ending motivations. I will be forever grateful to him for giving me the opportunity to be his student.

I would like to offer my special thanks to my dear uncle, and my best friends Aryan, Faraz, Mojtaba, Mohammad, Parvin, Mahmoud, Hossein, Reza, Amir, Mohsen, Ali, Ehsan, Edris, and Saeid who supported and helped me in different steps of my adventurous journey to Canada and starting my education

Sincere thanks to Mahdi, who is truly my brother, always ready to help everyone with his admirable knowledge, golden and giving heart.

During my studies, I had the chance to have Nuha beside of me and I would like to thank her for her assistance.

Special thanks to my knowledgeable, kind friend and lab mate, Walid, who always helped me out whenever I needed him. Also, I am very thankful to my lab mates Jaspreet, Kamal, Narges, Muhammad, Hieu and others.

It is noteworthy to mention my gratitude to Ms. Diane Moffat and Sheryl Tablan who supported me with their kind helps and advices, always having beautiful smiles.

Last but not the least, I would like to thank my mother and brother, as the most valuable gift of my life because of their unconditional love, respect and trust. They always encourage me to achieve my goals, believe in my dreams and standing on my own feet, which let me have confidence to discover new paths in my life. I am truly blessed to have you. Thank you for everything and whatever you have done for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

One of the main problems in data mining and machine learning, is the selection of a proper representation of data [56]. Most of the learning methods have been developed for the case of continuous data. However, in this work, we consider count data, which are naturally generated by a huge number of applications in several fields such as machine learning, computer vision, pattern recognition, and data mining (*e.g.* [7, 1, 17, 70]). For example, we could consider images or textual documents modeling and clustering, where each image or text can be represented by a vector of frequencies of visual words or words, respectively [8]. The Poisson distribution is a primary distribution for modeling count data, that has an equal mean and variance (equi-dispersion). However, in many practical situations this assumption is not valid as real data exhibits the phenomenon of overdispersion (*i.e.*, the variance of the count variable exceeds its mean) [53]. Consider for example, an image represented by the bag-of-features approach, where some of the features appear more frequently than others, or do not appear at all, making the variance greater than the mean. For addressing this issue, the negative-binomial distribution has been widely used in high-throughput sequencing data [30, 4]. Moreover, multinomial (MN) distribution is another fundamental model in count data analysis which is useful for analyzing count proportions between multiple categories. For instance, a comparison between the multinomial model and three probabilistic models: negative binomial, Bernoulli and Poisson, with the application to text classification, has shown that the multinomial usually performs better [24]. However, in real data we encounter another phenomenon caused by dependencies or the similarity of responses of members of the same cluster. This leads to extra-multinomial variation [29], *i.e.* overdispersion with respect to the MN distribution. Thus, the multinomial distribution has been extended to the Multinomial-Dirichlet Distribution (MDD) [48, 52, 14, 41, 11, 10] to model the overdispersion of the MN distribution. Having the Dirichlet as a conjugate prior to the multinomial has some computational advantages [42], which has made Dirichlet a popular choice as a prior in the case of contingency tables [3], Bayesian belief nets [31], and in the context of graphical models [26]. The MDD distribution has been used in various fields, including topic and magazine exposure modeling [46, 33, 58], word burstiness modeling [41], and language modeling [40]. On the other hand, having a highly restrictive negative covariance structure, makes the Dirichlet's usage as a prior improper, when we have positively correlated data. The multinomial generalized Dirichlet (MGDD) could be considered as an alternative to the MDD, that has a more general covariance structure [7]. However, a D-variate MGDD has 2D parameters versus D + 1 parameters for a D-variate MDD, which means requiring the learning of a greater number of parameters. Also, the author in [9] proposed another alternative that is the composition of the Beta-Liouville distribution and the multinomial, named multinomial Beta-Liouville distribution (MBLD), which has less parameters than the recently proposed MGDD model. In MBLD, the Beta-Liouville is selected from the Liouville family of distributions, including the Dirichlet as a special case, which is a conjugate prior to the multinomial, and also like generalized Dirichlet, it has a general covariance structure.

Given the importance of count data, there have been numerous efforts for analyzing this kind of data using both supervised and unsupervised learning approaches. Finite mixture models are among the most widely used techniques for data clustering [43, 62]. Generally, these models are based on the clustering (partitioning) of the data into a number of groups, where each group contains vectors that have a certain degree of similarity. In fact, many studies have proved that the adoption of discrete finite mixture models can have higher performance as compared with other common used approaches such as neural networks and decision trees [61]. Finite mixtures are popular for modeling univariate and multivariate data [45]. Novel machine learning applications have changed the direction of the current research activities from working on mixtures for continues data to other types of data such as binary or integer-valued features [50] applied in text classification, and binned and truncated multivariate data [16]. In the majority of the cases, the probability density functions (PDFs) of mixture models are considered to be Gaussian, which is not the best choice, specially where the partitions are clearly non-Gaussian [5]. For example, it has been shown that in the case of modeling discrete data in computer vision, Gaussian assumption is an inadequate choice, and most of the researchers use the multinomial distribution [50, 63]. A few works however, have been proposed for count data clustering, especially when dealing with textual documents [63]. Flexibility and offering a systematic formal way to unsupervised learning, make mixture models a popular method in this field [6].

1.2 Objectives

The main objective of this thesis is to introduce an overdispersed count data clustering algorithm, which uses a novel technique for computing the log-likelihood function, of the multinomial Dirichlet, multinomial generalized Dirichlet, and multinomial Beta-Liouville distributions. We developed a learning framework based on maximum likelihood estimation to obtain optimal parameters for our proposed mixture models and applied it to address the following challenging issues:

1. Selecting a flexible mixture model with higher efficiency (clustering accuracy) in modeling overdispersed multicategorical count data, in compare with traditional methods.
2. Parameters estimation as one of the critical and crucial challenges when deploying mixture models.
3. Evaluating the performance and feasibility of the proposed approach, through a set of empirical experiments involving real world applications.
4. Comparison of the efficiency of our framework using three different distributions.

For fitting the parameters of the mixture models to our observed data (learning the model), we use the expectation-maximization (EM) algorithm, for determining a maximum likelihood (ML) estimation of the mixture parameters [23]. Since the log likelihood function has an essential task in such statistical inference method [18, 49], some authors such as Yu and Shaw [69] proposed a novel parameterization of the MDD log-likelihood function based on truncated series consisting of Bernoulli polynomials, and for expanding its applicability, they extend it to a mesh algorithm for computing the log-likelihood. In this work, we first adopted this mesh algorithm for the computation of the MDD log-likelihood within a mixture model framework. Afterwards, we extended the approach in [69] to reparameterize the MGDD and MBLD log-likelihood functions, along with utilizing the mesh algorithm for the computation of the MGDD and MBLD loglikelihoods into a mixture model framework and parameter estimation process. We evaluated our clustering approach on various real-world challenging problems.

1.3 Contributions

Our major contributions in this thesis are as follows:

1. Proposing a novel finite mixture model for overdispersed multicategorical count data with the MDD distribution. Our proposed framework is based on the approximation of the paired log-gamma difference of the loglikelihood function, and expanding its applicability by using the mesh algorithm, which to the extent of our knowledge it has not been used before for clustering. We have proven that using this method could result to higher clustering accuracy as compared with prevalent technique. This contribution has been published in *Mixture Models and Applications, Unsupervised and Semi-supervised Learning, Springer Nature*, as a book chapter.

2. Extending the previous approach to reparameterize the MGDD and MBLD log-likelihood functions, and also applying the mesh algorithm for the computation of the MGDD and MBLD loglikelihood into a finite mixture model framework. Part of this contribution has been published in the above mentioned book as a chapter, and the other part has been submitted to *IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2019).*

3. Investigating the performance of our framework by testing it on challenging real problems such as natural scenes categorization and facial expression recognition.

1.4 Thesis Overview

In this thesis, we propose an overdispersed count data clustering algorithm in the following chapters:

1. In chapter 2, we present our modeling approach. First, the MDD distribution is presented, and the parameterization of the MDD log-likelihood function is reviewed, to allow smooth transition from the overdispersed to the non-overdispersed case. Then, we discusses the MGDD and MBLD distributions and propose their new parameterizations, and also describe the mesh algorithm for computing the log-likelihood functions. We end this chapter by discussing clustering, using finite mixture models.

2. Chapter 3 is devoted to the experimental results. We proved the performance of our work experimentally via the challenging natural scenes categorization and facial expression recognition applications and analyzing the results.

3. Finally in chapter 4, we conclude our work, highlight some challenges and suggest future works.

Chapter 2

Computation of Log-likelihood Functions

In this chapter, we discuss the computation of the log-likelihood functions of the MDD, MGD, and MBL distributions.

2.1 Computing the MDD Log-likelihood Function

In this section, we first discuss the Multinomial Dirichlet Distribution (MDD) in details. Later on, the approximation of the paired log-gamma difference of the log-likelihood function is explained.

2.1.1 The Multinomial Dirichlet Distribution

We assume that $O = (O_1, ..., O_N)$ is a finite set of abstract objects and $e = (e_1, ..., e_K)$ the domain of some events. Also, we consider that the counts for each object O_i are available as a co-occurrence vector $\vec{X}_i = (X_{i1}, ..., X_{iK})$, where X_{ij} refers to the number of times events e_j happens in the object O_i. Hence, we represent the object by $\vec{X}$ supposing that $\vec{X}$ follows a Multinomial distribution. However, using frequencies for obtaining the probabilities gives a weak estimation, due to the fact that the events are considered independent, which is not always true [19, 32]. Several attempts have been made to address this issue. Teevan and Karger developed a model that fits discrete vectors in an exponential family of models [60]. Rennie et al. tried to reduce the impact of dependency by log-normalizing the counts [55].

Consider the observations (vector of counts) $\vec{X} = (X_1, ..., X_K)$, satisfying $\sum_{k=1}^{K} X_k = N$, and $\vec{P} = (P_1, ..., P_K)$ satisfying $\sum_{k=1}^{K} P_k = 1$, where P_k is the probability of seeing the kth feature. The probability mass function (PMF) of K categories of the MN distribution having N-independent trials is given by:

$$\mathcal{M}(\vec{X}|\vec{P}) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} P_k^{X_k} \tag{1}$$

The Dirichlet distribution is a conjugate prior of $\vec{P}$. Suppose the random vector $\vec{P} = (P_1, ..., P_K)$ follows a Dirichlet distribution with parameters $\alpha = (\alpha_1, ..., \alpha_K)$, the joint density function is [36]:

$$\mathcal{D}(P_1, ..., P_k) = \frac{\Gamma(A)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} P_k^{(\alpha_k - 1)} \tag{2}$$

where $\Gamma(\alpha_k)$ is the gamma function and $A = \sum_{k=1}^{K} \alpha_k$.

The marginal distribution of $\vec{X}$ is obtained by taking the integral of the product of the Dirichlet prior and the Multinomial likelihood, with respect to the probabilities $\vec{P}$ [48]:

$$\mathcal{MDD}(\vec{X}|\vec{\alpha}) = \frac{N!}{\prod_{k=1}^{K} X_k!} \frac{\Gamma(A)}{\Gamma(A+N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + X_k)}{\Gamma(\alpha_k)} \tag{3}$$

We call this density the MDD (Multinomial Dirichlet distribution) and the mean and variance of this distribution are given by [48]:

$$E(X_i) = \frac{|\vec{X}|\alpha_i}{\vec{\alpha}} \tag{4}$$

$$Var(X_i) = \frac{|\vec{X}|(|\vec{X}|-1)\alpha_i(|\alpha|-\alpha_i)}{(|\alpha|)^2(|\alpha|+1)} + \frac{|\vec{X}|\alpha_i}{\vec{\alpha}} \tag{5}$$

A special case of the MDD distribution with just two-parameters ($K = 2$) is named the Beta-Binomial distribution, that has been widely studied by [27, 38]. The first term on the right side of Eq.(3) does not depend on the parameters α. For the maximum-likelihood estimation, we are not interested in the first term but in the product of the remaining two terms of the MDD likelihood function in Eq.(3). By taking the logarithm of both sides of the above equation, we achieve the log-likelihood function:

$$\begin{aligned}\ln \mathcal{L}(P,\psi;X) = &-(\ln\Gamma(1/\psi+N) - \ln\Gamma(1/\psi)) \\ &+\sum_{k=1}^{K}\Bigg(\ln\Gamma\Bigg(1/\frac{\psi}{P_k}+X_k\Bigg) - \ln\Gamma\Bigg(1/\frac{\psi}{p_k})\Bigg)\Bigg)\end{aligned} \tag{6}$$

where $\psi = 1/A$ and $p = \psi\alpha$. In this work, we call ψ the overdispersion parameter, which has a direct relation with the variance, and specifies the difference between a MDD distribution and its corresponding MN distribution in the same probability category. This formula has some deficiencies including that it is undefined for $\psi = 0$, also it is unstable when $\psi \to 0$, since each $\ln\Gamma$ term becomes very large, and the paired differences become relatively small which result in computation errors. The mesh algorithm, proposed in [69], applies a new formula based on a truncated series consisting of Bernoulli polynomials to solve the instability problem without incurring long run times.

2.1.2 Approximating the Paired Log-Gamma Difference in MDD Log-likelihood Function

As proposed in [69], we use the approximation of the paired log-gamma difference method and the properties of analytic functions as follows [57, 15]:

$$\ln\Gamma(1/x+y) - \ln\Gamma(1/x) \approx -y\ln x + D_m(x,y) \tag{7}$$

when y is an integer, $|x|\, min(|y-1|,|y|) < 1$, $xy \leq \delta$ and:

$$D_m(x,y) = \sum_{n=2}^{m}\frac{(-1)^n\phi_n(y)}{n(n-1)}x^{(n-1)} \tag{8}$$

where

$$\phi_n(y) = B_n(y) - B_n \tag{9}$$

is the old type Bernoulli polynomial [65], $B_n(y)$ and B_n indicate the nth Bernoulli polynomial, and nth Bernoulli number ($B_n = B_n(0)$), respectively.

Using floating-point arithmetic, high order polynomials are hard to compute [22]. Hence, we cannot use too large m's, because the error of each terms of $D_m(x,y)$

may be large, which eventually makes it inaccurate. Following Yu and Shaw [69], for computing the log-likelihood of the MDD distribution using the mesh method, we used $m = 20$, as it makes $\phi_n(y)$ $(n \leq m)$ numerically accurate. We also choose $\delta = 0.2$, that results to an error bound of $\sim 1.30 \times 10^{-16}$, which is just little less than the machine epsilon double precision data type $\approx 2.22 \times 10^{-16}$.

Let $\vec{X}^+$ be the vector of the non-zero elements in $\vec{X}$, $\vec{P}^+$ be a vector of the corresponding elements in $\vec{P}$, and K^+ be the length of $\vec{X}^+$, then Eq.(6) becomes:

$$\begin{aligned} \ln \mathcal{L}(P^+, \psi; X^+) = -\Big(& \underbrace{\ln \Gamma(1/\psi + N) - \ln \Gamma(1/\psi)}_{*} \Big) \\ & + \sum_{k=1}^{K^+} \Bigg(\underbrace{\ln \Gamma(1/\frac{\psi}{P_k^+} + X_k^+) - \ln \Gamma(1/\frac{\psi}{P_k^+})}_{**} \Bigg) \end{aligned} \tag{10}$$

As mentioned earlier, when condition $xy \leq \delta$ is met, we can use the approximation in Eq.(7) for all $K^+ + 1$ paired log-gamma differences in Eq. 10, as [69]:

$$\begin{aligned} \ln \mathcal{L}(P^+, \psi; X^+) &\approx -(-N \ln \psi + D_m(\psi, N)) + \sum_{k=1}^{K^+} \Bigg(- X_k^+ \ln \left(\frac{\psi}{P_k^+} \right) + D_m \left(\frac{\psi}{P_k^+}, X_k^+ \right) \Bigg) \\ &= -D_m(\psi, N) + \sum_{k=1}^{K^+} \Bigg(X_k^+ \ln p_k^+ + D_m \left(\frac{\psi}{P_k^+}, X_k^+ \right) \Bigg) \end{aligned} \tag{11}$$

2.2 Computing the MGDD Log-likelihood Function

In this section we discuss the MGD distribution in sufficient details. Then, we propose the approximation of the paired log-Gamma differences technique for computing the MGDD log-likelihood function.

2.2.1 The Multinomial Generalized Dirichlet Distribution

Despite its flexibility and its several interesting properties, such as the consistency of its estimates as a prior, the fact that it is conjugate to the multinomial, and its ease of use, the Dirichlet distribution has a very restrictive negative covariance structure that makes its use as a prior in the case of positively correlated data inappropriate. Another restriction of the Dirichlet distribution is that the variables with the same mean must have the same variance [35]. Recent works have shown that all these disadvantages can be handled by using the generalized Dirichlet distribution which has many convenient properties that make it more useful and practical, as a prior to the multinomial, than the Dirichlet in real-life applications [7, 67].

The generalized Dirichlet distribution with parameter vector $\vec{\alpha} = (\alpha_1, \beta_1, ..., \alpha_k, \beta_k)$ is defined by [20]:

$$\mathcal{G}(P_1, ..., P_K) = \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} P_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K} P_k\right)^{\gamma_k} \tag{12}$$

for $\sum_{k=1}^{K} P_k < 1, 0 < P_k < 1, k = 1, ..., K$, where $\alpha_k > 0$, $\beta_k > 0$, $\gamma_k = \beta_k - \alpha_{k+1} - \beta_{k+1}$, for $k = 1, ..., K-1$ and $\gamma_k = \beta_k - 1$.

Define $\vec{X} = (X_1, \ldots, X_{K+1})$ as a overdispersed vector of counts of $K+1$ events. Then, the composition of the generalized Dirichlet and the multinomial gives the Multinomial Generalized Dirichlet (MGD), as [7]:

$$\mathcal{MGD}(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(N+1)}{\prod_{k=1}^{K+1} \Gamma(X_{ik+1})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \prod_{k=1}^{K} \frac{\Gamma(\alpha'_k)\Gamma(\beta'_k)}{\Gamma(\alpha'_k + \beta'_k)}, \tag{13}$$

where $\alpha'_k = \alpha_k + X_{ik}$, and $\beta'_k = \beta_k + X_{ik+1} + \cdots + X_{iK+1}$ for $k = 1, \ldots, K$, and $\sum_{k=1}^{K+1} X_{ik} = N$. Given that the generalized Dirichlet includes the Dirichlet as a special case, MGDD is reduced to a MDD when $\beta_k = \alpha_{k+1} + \beta_{k+1}$.

The mean and the variance of the generalized Dirichlet distribution satisfy the following conditions [20, 66]:

$$E(P_k) = \frac{\alpha_k}{\alpha_k + \beta_k} \prod_{l=1}^{k-1} \frac{\beta_l}{\alpha_l + \beta_l}, \tag{14}$$

$$Var(P_k) = E(P_k)\left(\frac{\alpha_k + 1}{\alpha_k + \beta_k + 1} \prod_{l=1}^{k-1} \frac{\beta_l + 1}{\alpha_l + \beta_l + 1} - E(P_k)\right), \tag{15}$$

and the covariance between P_{k1} and P_{k2} is:

$$Cov(P_{k_1}, P_{k_2}) = E(P_{k_2})\left(\frac{\alpha_{k_1}}{\alpha_{k_1} + \beta_{k_1} + 1} \prod_{l=1}^{k_1 - 1} \frac{\beta_l + 1}{\alpha_l + \beta_l + 1} - E(P_{k_1})\right) \tag{16}$$

Like the Dirichlet, the generalized Dirichlet is conjugate to the multinomial distribution but has a more general covariance structure than the Dirichlet distribution and the variables with the same mean do not need to have the same variance [13, 7], thus, it is more practical to be used in modelling data with overdispresion. Moreover, it remains K degrees of freedom, which makes it more flexible for several real-world applications [34].

2.2.2 Approximating the Paired Log-Gamma Difference in MGDD Log-likelihood Function

The first term on the right side of Eq.(13) does not depend on the parameters α and β. For the maximum-likelihood estimation, we are not interested in the first term but in the product of the remaining two terms of the MGDD likelihood function in Eq.(13):

$$\mathcal{L}(\alpha, \beta; X) = \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \prod_{k=1}^{K} \frac{\Gamma(\alpha'_k)\Gamma(\beta'_k)}{\Gamma(\alpha'_k + \beta'_k)} \tag{17}$$

By taking the logarithm of both sides of Eq.(17), we get the log-likelihood function, as:

$$\begin{aligned} \ln \mathcal{L}(\alpha, \beta; X) = & \sum_{k=1}^{K} (\ln \Gamma(\alpha_k + \beta_k) - \ln \Gamma(\alpha_k) - \ln \Gamma(\beta_k)) \\ & + \sum_{k=1}^{K} - (\ln \Gamma(\alpha'_k + \beta'_k) - \ln \Gamma(\alpha'_k) - \ln \Gamma(\beta'_k)) \end{aligned} \tag{18}$$

Similar to the case of MDD, we consider the parameter ψ as the overdispersion parameter that gives the MGD distribution the ability to capture data variation. Using $\psi = 1/A$ where $A = \Sigma_{k=1}^{K} \alpha_k$ and $P = \psi\alpha$, thus, Eq. (18) becomes:

$$\begin{aligned} \ln \mathcal{L}(P, \psi, \beta; X) = & \sum_{k=1}^{K} \left(\ln \Gamma(1/\frac{\psi}{P_k} + \beta_k) - \ln \Gamma(1/\frac{\psi}{P_k}) \right) \\ + & \sum_{k=1}^{K} - \left(\ln \Gamma(1/(\frac{\psi}{P_k + X_k\psi}) + \beta'_k) - \ln \Gamma(1/(\frac{\psi}{P_k + X_k\psi})) \right) \\ + & \sum_{k=1}^{K} \left(\ln \Gamma(1/\frac{1}{\beta_k} + (X_{ik+1} + \cdots + X_{iK+1})) - \ln \Gamma(\frac{1}{\beta_k}) \right) \end{aligned} \tag{19}$$

Considering $\vec{X}^{+}$, $\vec{P}^{+}$, $\vec{\beta}^{+}$, $\vec{\beta'}^{+}$ vectors of non-zero elements in $\vec{X}$, $\vec{P}$, $\vec{\beta}$ and $\vec{\beta}'$

respectively, where K^+ is the length of $\vec{X}^+$, then, Eq. (19) becomes:

$$
\begin{aligned}
\ln \mathcal{L}(P^+, \psi, \beta^+; X^+) = \sum_{k=1}^{K^+} \left(\underbrace{\ln \Gamma(1/\frac{\psi}{P_k^+} + \beta_k^+) - \ln \Gamma(1/\frac{\psi}{P_k^+})}_{*} \right) \\
+ \sum_{k=1}^{K^+} - \left(\underbrace{\ln \Gamma(1/(\frac{\psi}{P_k^+ + X_k^+ \psi}) + \beta_k'^+) - \ln \Gamma(1/(\frac{\psi}{P_k^+ + X_k^+ \psi}))}_{**} \right) \\
+ \sum_{k=1}^{K^+} \left(\underbrace{\ln \Gamma(1/\frac{1}{\beta_k^+} + (X_{ik+1}^+ + \cdots + X_{iK+1}^+)) - \ln \Gamma(\frac{1}{\beta_k^+})}_{***} \right)
\end{aligned} \tag{20}
$$

Similar to the approach in [69], if the condition $xy \leq \delta$ is met, we can use the approximation (7) for all $K^+(^*)$, $K^+(^{**})$, and $K^+(^{***})$ paired log-gamma differences in (20):

$$
\begin{aligned}
\ln \mathcal{L}(P^+, \psi, \beta^+; X^+) = \sum_{k=1}^{K^+} \left(-\beta_k^+ \ln(\frac{\psi}{P_k^+}) + D_m(\frac{\psi}{P_k^+}, \beta_k^+) \right) \\
+ \sum_{k=1}^{K^+} \left(\beta_k'^+ \ln(\frac{\psi}{P_k^+ + X_k^+ \psi}) - D_m(\frac{\psi}{P_k^+ + X_k^+ \psi}, \beta_k'^+) \right) \\
+ \sum_{k=1}^{K^+} \left(-(X_{ik+1}^+ + \cdots + X_{iK+1}^+) \ln(\frac{1}{\beta_k^+}) + D_m(\frac{1}{\beta_k^+}, (X_{ik+1}^+ + \cdots + X_{iK+1}^+)) \right)
\end{aligned} \tag{21}
$$

Here, we also used the same values for $m = 20$ and $\delta = 0.2$ used for MDD .

2.3 Computing the MBLD Log-likelihood Function

In this section we discuss the MBL distribution in adequate details. Then, we propose the approximation of the paired log-gamma differences method for computing the MBLD log-likelihood function.

2.3.1 The Multinomial Beta-Liouville Distribution

We assume that we have N observations (objects) $\mathcal{X} = (\vec{X_1}, ..., \vec{X_N})$, where each vector $\vec{X_i}$ could for example represent a given image or document, and is defined as an overdispersed vector of counts (or frequencies) of $K+1$ features, $\vec{X} = (X_1, ..., X_{K+1})$, satisfying $\sum_{k=1}^{K+1} X_k = N$, and $\vec{P} = (P_1, ..., P_{K+1})$ satisfying $\sum_{k=1}^{K+1} P_k = 1$, where P_k being the probability of observing the kth feature. The features could be considered as visual words if we have images, or words in the case of textual documents. Since $\vec{X}$ is a vector of counts, a popular assumption to make is that it follows a multinomial distribution. The probability density function (PDF) of K+1 categories with N-independent trials is given by:

$$\mathcal{M}(\vec{X}|\vec{P}) = \frac{N!}{\prod_{k=1}^{K+1} X_k!} \prod_{k=1}^{K+1} P_k^{X_k} \tag{22}$$

In which as we can see, frequencies are used to set the probabilities. However, as we mentioned earlier, using only frequencies is one of its limitations, since events are considered independent and also we are assigning low probabilities to infrequent or even unseen events, especially when the data are overdispersed and sparse [19, 32], which is exactly our case study. As we discussed before, different solutions have been sugested to solve this issue [55, 60]. Using prior information into the construction of the statistical model is particularly an attempt, which in this section, it is chosen to be the Beta-Liouville distribution.

Suppose the vector $\vec{P} = (P_1, ..., P_K)$ follows a Beta-Liouville distribution, with parameters $\theta = (\alpha_1, ..., \alpha_K, \alpha, \beta)$, where $u = \sum_{k=1}^{K} P_k < 1, P_k > 0, k = 1, ..., K$. Then:

$$\mathcal{B}(P_1, ..., P_k) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times u^{\alpha - \sum_{k=1}^{K} \alpha_k}(1-u)^{\beta-1} \prod_{k=1}^{K} \frac{P_k^{\alpha_k - 1}}{\Gamma(\alpha_k)} \tag{23}$$

is called the Beta-Liouville distribution [2, 12]. We obtain the mean, variance, and the covariance of this ditribution by replacing (24) and (25) into (26), (27), and (28) respectively, as follows:

$$E(u) = \frac{\alpha}{\alpha+\beta} \tag{24}$$

$$E(u^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \tag{25}$$

$$E(P_k) = E(u)\frac{\alpha_k}{A} \tag{26}$$

$$Var(P_k) = E(u^2)\frac{\alpha_k(\alpha_k+1)}{A(A+1)} - E(P_k)^2\frac{\alpha_k^2}{A^2} \tag{27}$$

$$Cov(P_{k_1}, P_{k_2}) = \frac{\alpha_{k_1}\alpha_{k_2}}{\sum_{k_1=1}^{K}\alpha_{k_1}}\left(\frac{E(u^2)}{A+1} - \frac{E(u)^2}{A}\right) \tag{28}$$

where A $= \sum_{k=1}^{K}\alpha_k$.

We may obtain the marginal distribution of $\vec{X}_i$ by integrating over the product of the prior Beta-Liouville and the multinomial likelihood, with respect to $\vec{P}$:

$$\begin{aligned}\mathcal{MBD}(\vec{X}_i|\theta) = {} & \frac{\Gamma(N+1)}{\prod_{k=1}^{K+1}\Gamma(X_{ik+1})} \\ & \times \frac{\Gamma(A)\Gamma(\alpha+\beta)\Gamma(\alpha')\Gamma(\beta')\prod_{k=1}^{K}\Gamma(\alpha'_k)}{\Gamma(A')\Gamma(\alpha'+\beta')\Gamma(\alpha)\Gamma(\beta)\prod_{k=1}^{K}\Gamma(\alpha_k)}\end{aligned} \tag{29}$$

where $\alpha'_k = \alpha_k + X_{ik}$, $\alpha' = \alpha + \sum_{k=1}^{K} X_{ik}$, $\beta' = \beta + X_{iK+1}$, $\sum_{k=1}^{K+1} X_{ik} = N$, and $A' = \sum_{k=1}^{K}\alpha'_k$.

This density is called the MBLD [9] which contains D + 2 parameters. There are interesting properties of the Liouville distribution which can be found in [59, 28].

The first term on the right side of Eq.(29), does not depend on the θ parameters. For estimation of maximum-likelihood, we are not interested in the first term, but in the remaining term that includes the distribution parameters. By taking the logarithm of both sides of Eq.(29), we obtain the log-likelihood function:

$$
\begin{aligned}
&\ln \mathcal{L}(\theta; X) \\
&= -(\ln\Gamma(A') - \ln\Gamma(A)) \\
&- (\ln\Gamma(\alpha' + \beta') - \ln\Gamma(\alpha + \beta)) \\
&+ (\ln\Gamma(\alpha') - \ln\Gamma(\alpha)) + (\ln\Gamma(\beta') - \ln\Gamma(\beta)) \\
&+ (\ln(\prod_{k=1}^{K}\Gamma(\alpha'_k)) - \ln(\prod_{k=1}^{K}\Gamma(\alpha_k)))
\end{aligned} \tag{30}
$$

As mentioned in the previous sections, ψ is considered to be the overdispersion parameter, which determines the difference between a MBL distribution and its corresponding MN distribution, in the same probability category, and gives the MBL distribution the ability to capture data variation where $\psi = 1/A$ and $p = \psi\alpha$. Hence, Eq.(30) becomes:

$$
\begin{aligned}
&\ln \mathcal{L}(P, \psi, \alpha, \beta; X) \\
&= -\Big(\ln\Gamma(1/\psi + N') - \ln\Gamma(1/\psi)\Big) \\
&- \Big(\ln\Gamma(1/\frac{1}{\alpha+\beta} + N) - \ln\Gamma(1/\frac{1}{\alpha+\beta})\Big) \\
&+ \Big(\ln\Gamma(1/\frac{1}{\alpha} + N') - \ln\Gamma(1/\frac{1}{\alpha})\Big) \\
&+ \Big(\ln\Gamma(1/\frac{1}{\beta} + X_{iK+1}) - \ln\Gamma(1/\frac{1}{\beta})\Big) \\
&+ \sum_{k=1}^{K}\Big(\ln\Gamma(1/\frac{\psi}{P_k} + X_{ik}) - \ln\Gamma(1/\frac{\psi}{P_k})\Big)
\end{aligned} \tag{31}
$$

where $\sum_{k=1}^{K} X_{ik} = N'$.

One of the shortages of Eq.(31) is that it is undefined for $\psi = 0$, also unstable when $\psi \to 0$, because it makes each of the $\ln\Gamma$ terms very large, that causes the paired differences becoming relatively small, which leads to computation errors. As mentioned earlier, we are adopting the mesh algorithm, proposed by the authors in [69], which strives to address the instability problem without resulting in long run times.

2.3.2 Approximating the Paired Log-Gamma Difference in MBLD Log-likelihood Function

Considering $\vec{X}^{+}$ and $\vec{P}^{+}$, vectors of non-zero elements in $\vec{X}$ and $\vec{P}$, also α^{+} and β^{+} non-zero values, where K^{+} is the length of $\vec{X}^{+}$, then, Eq.(31) becomes:

$$
\begin{aligned}
&\ln \mathcal{L}(P^{+}, \psi, \alpha^{+}, \beta^{+}; X^{+}) \\
&= -\Big(\ln \Gamma(1/\psi + N') - \ln \Gamma(1/\psi) \Big) \\
&- \Big(\ln \Gamma(1/\frac{1}{\alpha^{+}+\beta^{+}} + N) - \ln \Gamma(1/\frac{1}{\alpha^{+}+\beta^{+}}) \Big) \\
&+ \Big(\ln \Gamma(1/\frac{1}{\alpha^{+}} + N') - \ln \Gamma(1/\frac{1}{\alpha^{+}}) \Big) \\
&+ \Big(\ln \Gamma(1/\frac{1}{\beta^{+}} + X^{+}_{iK+1}) - \ln \Gamma(1/\frac{1}{\beta^{+}}) \Big) \\
&+ \underbrace{\sum_{k=1}^{K^{+}} \Big(\ln \Gamma(1/\frac{\psi}{P^{+}_{k}} + X^{+}_{ik}) - \ln \Gamma(1/\frac{\psi}{P^{+}_{k}}) \Big)}_{*}
\end{aligned}
\tag{32}
$$

As discussed earlier, when the condition $xy \leq \delta$ is satisfied, we can use the approximation (7) for all $K^{+}(^{*}) + 4$ paired log-gamma differences in Eq.(32) as [69]:

$$
\begin{aligned}
&\ln \mathcal{L}(P^{+}, \psi, \alpha^{+}, \beta^{+}; X^{+}) \\
&= \Big(N' \ln(\psi) - D_m(\psi, N') \Big) \\
&+ \Big(N \ln(\frac{1}{\alpha^{+}+\beta^{+}}) - D_m(\frac{1}{\alpha^{+}+\beta^{+}}, N) \Big) \\
&+ \Big(-N' \ln(\frac{1}{\alpha^{+}}) + D_m(\frac{1}{\alpha^{+}}, N') \Big) \\
&+ \Big(-X^{+}_{iK+1} \ln(\frac{1}{\beta^{+}}) + D_m(\frac{1}{\beta^{+}}, X^{+}_{iK+1}) \Big) \\
&+ \sum_{k=1}^{K^{+}} \Big(-X^{+}_{ik} \ln(\frac{\psi}{P^{+}_{k}}) + D_m(\frac{\psi}{P^{+}_{k}}, X^{+}_{ik}) \Big)
\end{aligned}
\tag{33}
$$

After several attempts, the most optimum choice in our implementations was $m = 18$, which in our case made $\phi_n(y)$ $(n \leq m)$ numerically accurate. Here, we also used $\delta = 0.2$.

2.4 The Mesh Algorithm for Computing the Log-Likelihood Function

As discussed earlier, when condition $xy \leq \delta$ is met, it is possible to make use of the approximation in Eq. (7) for optimizing the log-gamma difference in computing the MDD, MGDD, and MBLD log-likelihood functions. However, when some of the terms in Eq.(10), (20), and (32) do not meet this condition; we may use the mesh algorithm, in which we rewrite the vector $\vec{X}$ into a sum of L terms, choosing the terms to meet the following condition:

$$X = \sum_{l=1}^{L} X^{(l)} \tag{34}$$

Afterwards, we specify the choice of $\alpha^{(l)}$, $\beta^{(l)}$, ${\alpha_k}^{(l)}$, and ${\beta_k}^{(l)}$ as below:

$$\alpha^{(l)} = \alpha + \sum_{i=1}^{l} X^{(i)}, \quad \text{for } l = 0, \ldots, L \tag{35}$$

$$\beta^{(l)} = \beta + \sum_{i=1}^{l} X^{(i)}, \quad \text{for } l = 0, \ldots, L \tag{36}$$

$${\alpha_k}^{(l)} = \alpha_k + \sum_{i=1}^{l} {X_k}^{(i)}, \quad \text{for } l = 0, \ldots, L \tag{37}$$

$${\beta_k}^{(l)} = \beta_k + \sum_{i=1}^{l} {X_k}^{(i)}, \quad \text{for } l = 0, \ldots, L \tag{38}$$

In other words, the adjacent ${\alpha_k}^{(l)}$'s have the following relations:

$${\alpha_k}^{(l-1)} + {X_k}^{(l)} = {\alpha_k}^{(l)}, \quad \text{for } l = 1, \ldots, L \tag{39}$$

or

$$P^{(l-1)}/\psi^{(l-1)} + X^{(l)} = P^{(l)}/\psi^{(l)}, \quad for\ l = 1, \ldots, L \tag{40}$$

Furthermore, given that $N^{(l)} = \sum_{k=1}^{K} X_k^{(l)}$ we have:

$$\frac{1}{\psi^{(l)}} = \frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}, \quad \text{for } l = 0, \ldots, L \tag{41}$$

$$\frac{1}{\psi^{(l)}} = \frac{1}{\psi^{(l-1)}} + N^{(l)}, \quad \text{for } l = 1, \ldots, L \tag{42}$$

The following relations also exist for all $\psi \in [0, +\infty]$:

$$\psi^{(l)} = \begin{cases} \dfrac{1}{\frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}} & \text{if } \psi \geq 1 \\ \\ \dfrac{\psi}{1 + \psi \sum_{i=1}^{l} N^{(i)}} & \text{if } 0 \leq \psi < 1 \end{cases} \tag{43}$$

$$P^{(l)} = \begin{cases} \dfrac{\frac{p}{\psi} + \sum_{i=1}^{l} X^{(i)}}{\frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}} & \text{if } \psi \geq 1 \\ \\ \dfrac{p + \psi \sum_{i=1}^{1} X^{(i)}}{1 + \psi \sum_{i=1}^{1} N^{(i)}} & \text{if } 0 \leq \psi < 1 \end{cases} \tag{44}$$

Therefore, for evaluating the log-likelihood function for MDD, MGDD, and MBLD, respectively, we use:

$$\ln \mathcal{L}(P^{+}, \psi; X^{+}) = \sum_{l=1}^{L} \ln \mathcal{L}(P^{(l-1)+}, \psi^{(l-1)}; X^{(l)+}) \tag{45}$$

$$\ln \mathcal{L}(P^{+}, \psi, \beta^{+}; X^{+}) = \sum_{l=1}^{L} \ln \mathcal{L}(P^{(l-1)+}, \psi^{(l-1)}, \beta^{(l-1)+}; X^{(l)+}) \tag{46}$$

$$\ln \mathcal{L}(P^{+}, \psi, \alpha^{+}, \beta^{+}; X^{+}) = \sum_{l=1}^{L} \ln \mathcal{L}(P^{(l-1)+}, \psi^{(l-1)}, \alpha^{(l-1)+}, \beta^{(l-1)+}; X^{(l)+}) \tag{47}$$

Each of the L terms in the above formulas, can be computed using Eq.(11) for MDD, Eq.(21) for the MGDD, and Eq.(33) for MBLD. This method is called the mesh algorithm [69], since the log-likelihood of the functions (45), (46), and (47) can be evaluated incrementally on a mesh. The mesh algorithm for computing MDD, MGDD and the MBLD log-likelihood functions can be described as follows:

1. For generating the mesh, we first create a primary mesh, using the following equation:
$$X_k^{(l)} = \lfloor \alpha_k^{(l-1)} \delta \rfloor \tag{48}$$

2. In the next step, the smallest integer meeting the following condition will be chosen as the mesh level L:
$$\sum_{l=1}^{L} X_k^{(l)} \geq X_k, \quad \text{for } k = 1, \ldots, K \tag{49}$$

   Generally, by adjusting the mesh level, we can properly select $X^{(l)}$, that all the terms in Eq.(45), (46), and (47) will meet the $xy \leq \delta$ condition. After many experiments, we selected $L = 3$ for MDD, $L = 5$ for MGDD, and $L = 6$ for MBLD, which has been found to be an appropriate choice as a mesh level.

3. Then, we adjust $X_k^{L'_k}$ so that $\sum_{l=1}^{L'_k} X_k^{(l)} = X_k$, and we will set all the remaining $X_k^{(l)}(l > L'_k)$ to zero, which means, the end of the mesh totals should exactly match X_k, considering L'_k being the smallest number when:
$$\sum_{l=1}^{L'_k} X_k^{(l)} \geq X_k, \quad \text{for } k = 1, \ldots, K \tag{50}$$

4. At the last stage, we implement Eq.(11), Eq.(21), and Eq.(33) for the computation of the MDD, MGDD, and MBLD log-likelihood functions, respectively.

2.5 Finite Mixture Model Learning

One of the most common methods for clustering is finite mixture modeling. In general, a collection of vectors contain objects, belonging to several clusters, modeled by a finite mixture of distributions. The principal goal here is to identify topological structure in a set of objects represented by count vectors, that is, to cluster objects that are similar enough to each other that might form different subcategories (i.e., components) of the objects. We can define a finite mixture model with M components as:

$$P(\vec{X}|\Theta) = \sum_{j=1}^{M} P(\vec{X}|j;\vec{\theta_j})P(j) \tag{51}$$

where symbol Θ is the entire set of parameters that needs to be be estimated: $\Theta = (\vec{\theta_1}, ..., \vec{\theta_M}, P(1), ..., P(M))$, $\vec{\theta_j}$ is the parameter vector for the jth component, $P(j)$ are the mixing proportions, satisfying $0 < P(j) \leq 1$ and $\sum_{j=1}^{M} P(j) = 1$. The maximum likelihood (ML) technique, has been the most popular method for estimating the parameters which determine a mixture, within the last two decades [54]. For estimating Θ using maximum likelihood, considering a set of N independent vectors $\mathcal{X} = (\vec{X_1}, ..., \vec{X_N})$, we have:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} P(\mathcal{X}|\Theta) \tag{52}$$

where:

$$P(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{M} P(\vec{X_i}|j;\vec{\theta_j})P(j) \tag{53}$$

By taking the logarithm of both sides of Eq.(53), we may have:

$$\Phi(\mathcal{X}, \Theta) = \sum_{i=1}^{N} log\Big(\sum_{j=1}^{M} P(\vec{X_i}|j;\vec{\theta_j})P(j)\Big) \tag{54}$$

Then, we apply the expectation-maximization (EM) algorithm for learning the mixture parameters and fitting the parameters of our mixture model into our observed data. This algorithm, produces a series of models with non-decreasing log-likelihood [44], assuming $\vec{X}$ as incomplete data.

The EM algorithm includes the following two steps:

E-step: Estimates the posterior probabilities, as:

$$P^{(t)}(j|\vec{X_i};\vec{\theta_j}) = \frac{P^{(t-1)}(\vec{X_i}|j;\vec{\theta_j})P^{(t-1)}(j)}{\sum_{j=1}^{M} P^{(t-1)}(\vec{X_i}|j;\vec{\theta_j})P^{(t-1)}(j)} \tag{55}$$

M-step: Updates the parameters estimation according to:

$$\hat{\Theta}^{(t)} = \arg\max_{\Theta}, \Phi(\mathcal{X}, \Theta^{(t-1)}) \tag{56}$$

By setting the derivatives of the log-likelihood function to zero, we may evaluate the $\vec{\theta_j}$ parameters for our distribution. However, due to the presence of $\Gamma(.)$ terms, we could not obtain a closed form solution in the M-step. Therefore, we use an iterative gradient descent optimization method by computing the gradient of the MDD likelihood, along with two bounds in equations [41, 47]. For the MGDD and MBLD, we use the Newton-Raphson method to estimate its parameters as proposed in [7] and [9].

Our initialization algorithm is defined as bellow:

1. For each of the j clusters (components), randomly generate the vector of parameters $\vec{\theta_j}$, to avoid having computational complexity which some of the other methods may have, such as method of moment. Note that our experiments show that this choice of initialization yields better results than the compared algorithms.

2. Implement the K-means algorithm, for allocating each of the observations (data points) to one of the existing clusters, assuming that the current model is accurate.

3. Finally, estimate the mixing proportions $P(j)$ as following:

   $$P(j) = \frac{\textit{number of elements in cluster j}}{N}$$

   where N being the number of data points.

At the end, we summarize the EM algorithm for learning the finite mixture model parameters in Algorithm 1.

Algorithm 1: EM algorithm for learning the mixture models parameters

```
Input  : $K$-dimensional data set with $N$ vectors $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$,
         pre-determined number of clusters $M$
State Initialize the set of parameters $\Theta$, as mentioned in section 2.5.
repeat
    State {E-step}
    for $i \leftarrow 1$ to $N$ do
        for $j \leftarrow 1$ to $M$ do
            Compute the posterior probabilities according to Eq.(55), where
              $P^{(t-1)}(\vec{X}_i|j;\vec{\alpha_j})$ is computed using the mesh algorithm explained
              in section 2.4.
        end
    end
    State {M-step}
    for $j \leftarrow 1$ to $M$ do
        Update the model parameters according to Eq.(56)
    end
until convergence;
```

Chapter 3

Experimental Results

In this chapter, we validate the performance of the proposed approach in clustering count, multi-categorial and overdispersed data with the MDD, MGDD and MBLD distributions, via two different applications: natural scenes categorization and facial expression recognition. In each application, we compare the accuracy of clustering different datasets, using the normal method and the mesh algorithm for the log-likelihood calculation.

For pre-processing, we used the SIFT (Scale-Invariant Feature Transform) [37] for feature extraction, and the Bag-of-Features (BoF) [21] for representation. BoF is based on the frequency of visual words, provided from a visual vocabulary, which is obtained by the quantization (or histogramming) of local feature vectors, computed from a set of training images. All the $128D$ descriptors calculated by SIFT are bined into a collection of local features. Afterwards, K-means is used to cluster the extracted vectors to build the visual words vocabulary. Then, every image in the data sets was represented by a vector, indicating the number of a set of visual words, coming from the constructed visual vocabulary. Since we used the iterative EM scheme, the initial parameter values might affect the convergence and the overall outcome. Hence, we run each model over 100 times with different random initializations in order to have fair results.

3.1 Natural Scenes Categorization

Image clustering is one of the most crucial topics in computer vision. In this experiment, we investigate the mesh algorithm's performance by scene clustering, which is a challenging application in the sense that in the real-life environment, they could be captured in various positions, distances, colors. Moreover, high probability of mis-clustering could be caused because of the noises that come from the background surroundings, that might have similar features as our natural scene targets.

In our experiments, we considered three different scene image datasets: SUN, Oliva and Torralba and Fei-Fei and Perona. Each dataset was split with 80:20 ratio, to form the visual vocabulary and representation.

SUN dataset is a subset of the extensive Scene UNderstanding (SUN) database[1][68], that contains 899 categories and 130,519 images. We use 1,849 natural scenes belonging to six categories (458 coasts, 228 rivers, 231 forests, 247 field, 518 mountains, and 167 sky/clouds). The average size of the images is 720 x 480 (landscape format) or 480 x 720 (portrait format). Samples from the considered subset are shown in Fig. 1.

Oliva and Torralba (OT) dataset [51], contains 2,688 images clustered as eight categories: 360 coasts, 328 forest, 374 mountain, 410 open country, 260 highway, 308 inside of cities, 356 tall buildings, and 292 streets. The average size of each image is 250 x 250 pixels. The last dataset is **Fei-Fei and Perona (FP)** [25], which includes 13 categories, only available in gray scale. This data set consists of the 2,688 images (eight categories) of the OT data set plus: 241 suburb residence, 174 bedroom, 151 kitchen, 289 living room, and 216 office. The average size of each image is approximately 250 x 300 pixels. Examples of images from these datasets are given in Fig. 2.

Table 1 represents the average clustering accuracies, using both the normal and mesh approach. As we can see, there are considerable improvements when we implemented our clustering algorithm using the mesh method. Among the tested data sets, SUN had the highest accuracy of 86.02% and 90.99%, modeled by MDD, using normal and mesh methods, respectively; from which we can observe an improvement of

[1] https://groups.csail.mit.edu/vision/SUN/

4.97% in the results when we implemented the mesh algorithm. The highest clustering accuracy using the same dataset, modeled by MGDD and MBLD, using normal method were 88.82% and 90.68%, following 2.51% and 1.25% growth when implemented by the mesh algorithm. The OT and FP datasets also experiment 3.66% and 1.59%, 2.7% and 1.13%, and 1.14% and 1.21% increase in the clustering accuracies when applying the mesh method, modeled by MDD, MGDD, and MBLD mixture models, respectively.

Moreover, Fig.3, Fig.4, Fig.5, Fig.6, Fig.7 and Fig.8 represent the confusion matrices when modeling the SUN dataset, with the MDD, MGDD, and MBLD distributions, employing the normal and mesh algorithms, respectively. From these figures, we can see that the best clustered objects are *coast* and *mountain*, which their accuracies have increased by 1.8% and 2.4%, and 2.9% and 0.1%, when using the mesh algorithm, modeled by MDD and MGDD finite mixture models, respectively. The greatest clustering accuracy gain was for *river* by 10.6% for MDD and *forest* by 6.1% for MGDD. On the other hand, in the MBLD model the highest clustering accuracies where for *forest* (97.5%) and *mountain* (92.5%) objects, when implemented by the normal method, but they did not experience a growth in their clustering accuracy when we applied the mesh technique. The sky images' clustering accuracy however, enhanced by 6.9% at the time we utilized the mesh method, which was the highest increase. Furthermore, we can notice that the misclassification between coast and river is happened because of having some similar features. Likewise, for the other scenes with a considerable amount of incorrectly clustered images: mountain and sky, and forest and field. Notable that the accuracy precision of sky has not changed (86%) when modeled by the MGDD distribution, at the time we employed the mesh algorithm. However, we are also seeing considerable percentage of misclusterings that have been reduced when adopting the mesh algorithm. For example, in the case of modeling the SUN dataset by the MBLD model, 2.3% of the misclassification of field objects as a mountain has been reduced to 0.0% at the time of using the mesh method. Same as misclassification of sky as a mountain, which has been decreased from 10.3% to 3.4%. All in all, the discussed facts had lead to having a higher percentage of clustering accuracy, when applying the mesh approach.

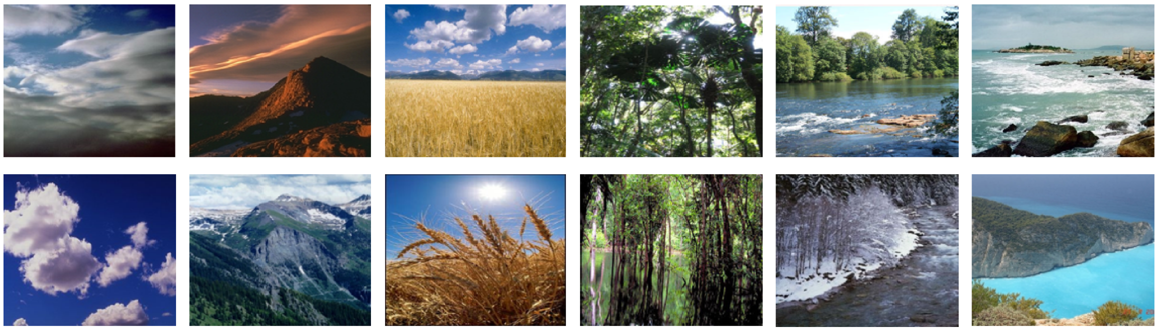


Figure 1: Sample images from the 6 categories in SUN dataset [68].

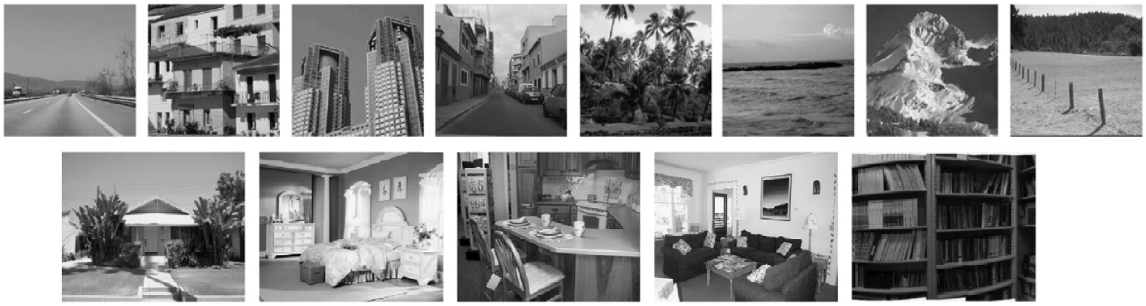
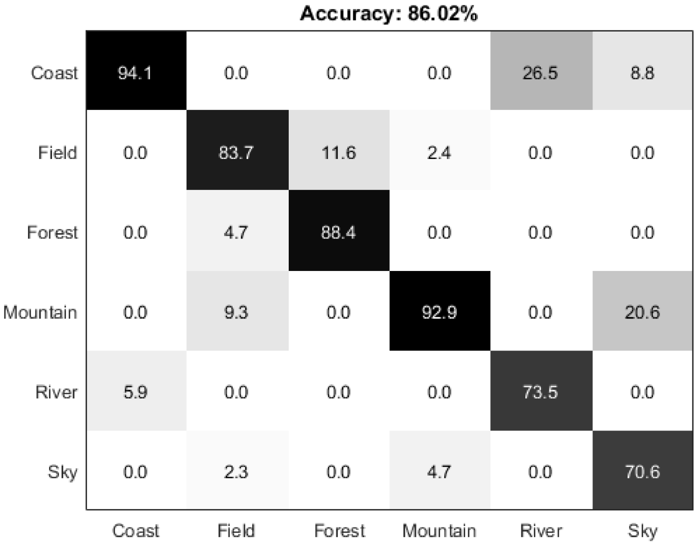


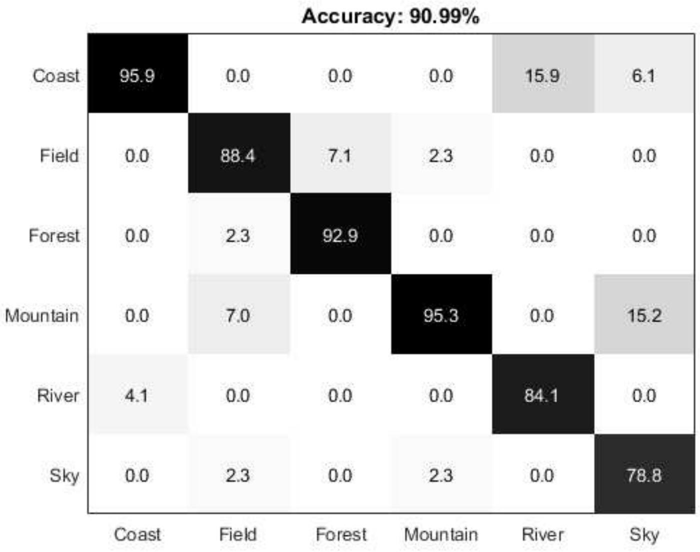
Figure 2: Example images from two of the used datasets. First row From OT [51], the second row contains the extra categories included in Fei-Fei and Perona dataset [25].



Accuracy: 86.02%

	Coast	Field	Forest	Mountain	River	Sky
Coast	94.1	0.0	0.0	0.0	26.5	8.8
Field	0.0	83.7	11.6	2.4	0.0	0.0
Forest	0.0	4.7	88.4	0.0	0.0	0.0
Mountain	0.0	9.3	0.0	92.9	0.0	20.6
River	5.9	0.0	0.0	0.0	73.5	0.0
Sky	0.0	2.3	0.0	4.7	0.0	70.6

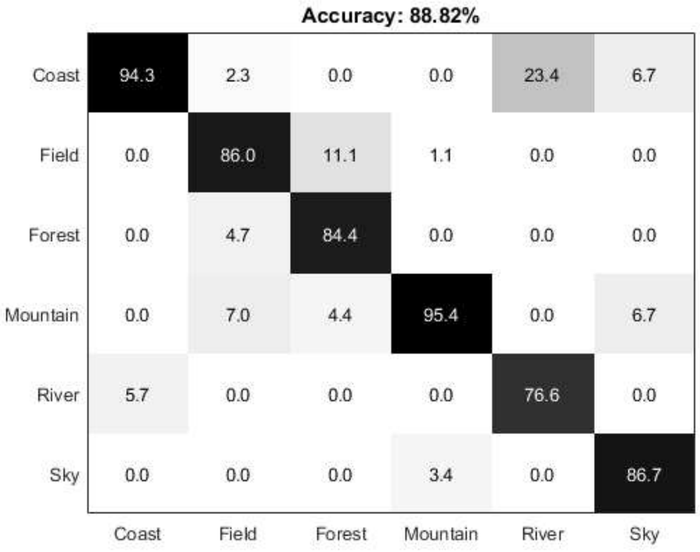
Figure 3: Confusion matrix for SUN dataset modeled by the MDD mixture, using normal method.



Accuracy: 90.99%

	Coast	Field	Forest	Mountain	River	Sky
Coast	95.9	0.0	0.0	0.0	15.9	6.1
Field	0.0	88.4	7.1	2.3	0.0	0.0
Forest	0.0	2.3	92.9	0.0	0.0	0.0
Mountain	0.0	7.0	0.0	95.3	0.0	15.2
River	4.1	0.0	0.0	0.0	84.1	0.0
Sky	0.0	2.3	0.0	2.3	0.0	78.8

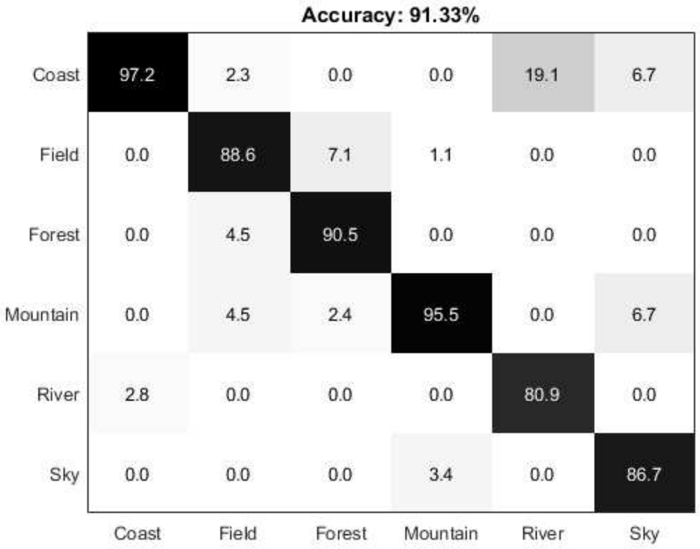
Figure 4: Confusion matrix for SUN dataset modeled by the MDD mixture, using mesh method.



Accuracy: 88.82%

	Coast	Field	Forest	Mountain	River	Sky
Coast	94.3	2.3	0.0	0.0	23.4	6.7
Field	0.0	86.0	11.1	1.1	0.0	0.0
Forest	0.0	4.7	84.4	0.0	0.0	0.0
Mountain	0.0	7.0	4.4	95.4	0.0	6.7
River	5.7	0.0	0.0	0.0	76.6	0.0
Sky	0.0	0.0	0.0	3.4	0.0	86.7

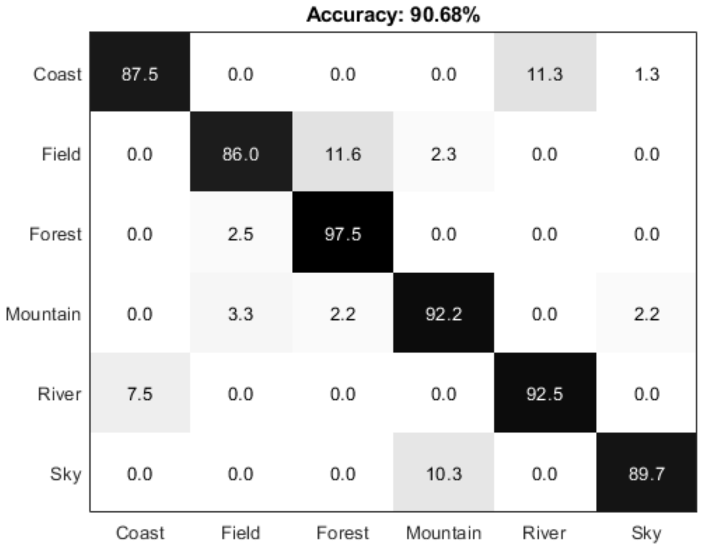
Figure 5: Confusion matrix for SUN dataset modeled by the MGDD mixture, using normal method.



Accuracy: 91.33%

	Coast	Field	Forest	Mountain	River	Sky
Coast	97.2	2.3	0.0	0.0	19.1	6.7
Field	0.0	88.6	7.1	1.1	0.0	0.0
Forest	0.0	4.5	90.5	0.0	0.0	0.0
Mountain	0.0	4.5	2.4	95.5	0.0	6.7
River	2.8	0.0	0.0	0.0	80.9	0.0
Sky	0.0	0.0	0.0	3.4	0.0	86.7

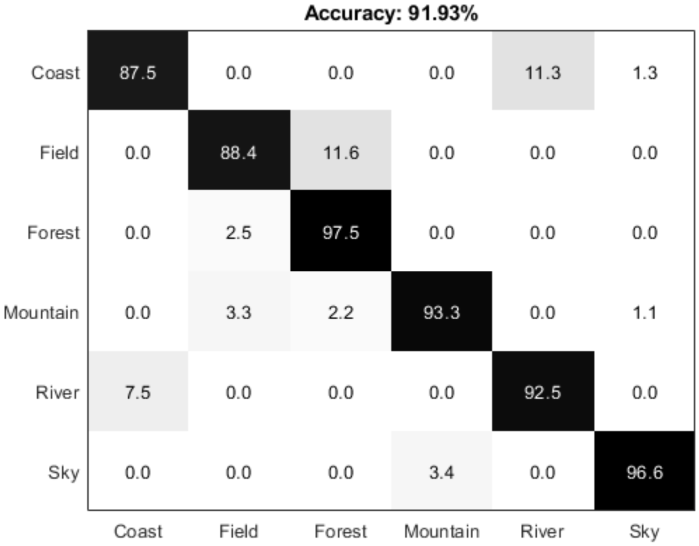
Figure 6: Confusion matrix for SUN dataset modeled by the MGDD mixture, using mesh method.



Accuracy: 90.68%

	Coast	Field	Forest	Mountain	River	Sky
Coast	87.5	0.0	0.0	0.0	11.3	1.3
Field	0.0	86.0	11.6	2.3	0.0	0.0
Forest	0.0	2.5	97.5	0.0	0.0	0.0
Mountain	0.0	3.3	2.2	92.2	0.0	2.2
River	7.5	0.0	0.0	0.0	92.5	0.0
Sky	0.0	0.0	0.0	10.3	0.0	89.7

Figure 7: Confusion matrix for SUN dataset modeled by the MBLD mixture, using normal method.



Accuracy: 91.93%

	Coast	Field	Forest	Mountain	River	Sky
Coast	87.5	0.0	0.0	0.0	11.3	1.3
Field	0.0	88.4	11.6	0.0	0.0	0.0
Forest	0.0	2.5	97.5	0.0	0.0	0.0
Mountain	0.0	3.3	2.2	93.3	0.0	1.1
River	7.5	0.0	0.0	0.0	92.5	0.0
Sky	0.0	0.0	0.0	3.4	0.0	96.6

Figure 8: Confusion matrix for SUN dataset modeled by the MBLD mixture, using mesh method.

3.2 Facial Expression Recognition

Facial expression recognition is one of the most important topics in various fields including computer vision and artificial intelligence. In fact, it is one of the most challenging tasks in social and interpersonal communication, since it is a natural way for human being to express emotions and therefore to show their intentions. The numerous number of expressions recognized in majority of the related databases makes the task hard as compared with other image categorization applications.

In this experiment, we used two different facial expression datasets: MMI and Extended Cohn-Kanade (CK+). This time, each dataset was split into two halves, to form the visual vocabulary and representation.

MMI [64] database includes 19 different faces of students and research staff of 300 members of both genders (44% female), ranging in age from 19 to 62, having either a European, Asian, or South American ethnic background. Currently it contains 2,894 image sequences where each image sequence has neutral face at the beginning and the end, and each with a the size of 720 x 576 pixels. We selected the sequences that could be labeled as one of the six basic emotions. Removing the natural faces results in 1,140 images.

The **Extended Cohn-Kanade (CK+)** [39] dataset consists of facial behavior of 210 adults 18 to 50 years of age. Image sequences were digitized into either 640 x 490 or 640 x 480 pixel arrays with 8-bit gray-scale value. We included all posed expressions that could be labeled as one of the 6 basic emotion categories which is about 4,000 images. (Table 2)

Examples from the used datasets are given in Fig. 9, Fig. 10.

Table 3 demonstrates clustering accuracy, using MDD, MGDD, and MBLD with normal and mesh log-likelihood calculation. By applying the mesh method, we are again having improvements when clustering both of the datasets as: 3.23% (MDD), 1.94% (MGDD), and 1.29% (MBLD) for MMI, and 2.88% (MDD), 1.72% (MGDD), and 0.72% (MBLD) for CK+. It is worth mentioning that, both of the applications, are experiencing better results (higher accuracy), when using normal and mesh methods, modeled by the MBLD mixture model.

Furthermore, figures 11, 12, and 13 depict the accuracy comparison of each cluster for the MMI dataset, modeled by MDD, MGDD, MBLD mixtures, respectively, using both normal and mesh methods. From Fig.11, it can be observed that the disgust, happiness and surprise emotions have gained 2.7%, 6.82% and 3.23% of accuracy respectively, when using the mesh algorithm. Moreover, Fig.12 shows that the happiness and sadness clusters are having 4.54% and 3.02% accuracy improvements, when mesh algorithm is implemented. In Fig.13 (MBLD), the happiness and sadness' accuracies have enhanced by the same percentage as in the Fig.12 (MGDD), but the surprise cluster have experimented 3.22% increase in clustering accuracy. Notable that the fear cluster's clustering accuracy did not change when modeled by the MDD and MGDD mixture models, applying both normal and mesh techniques. However, it had 100% correctness at the time it was modeled by the MBLD, using both of the methods. As we can see, the clustering accuracy of MBLD model is higher than MDD and MGDD, when employing both the normal and mesh approaches.

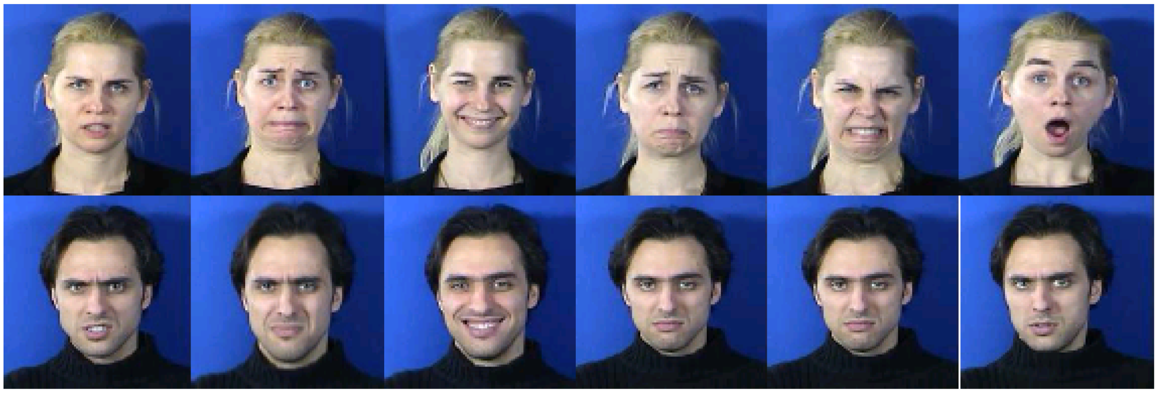


Figure 9: Example images from MMI dataset [64].

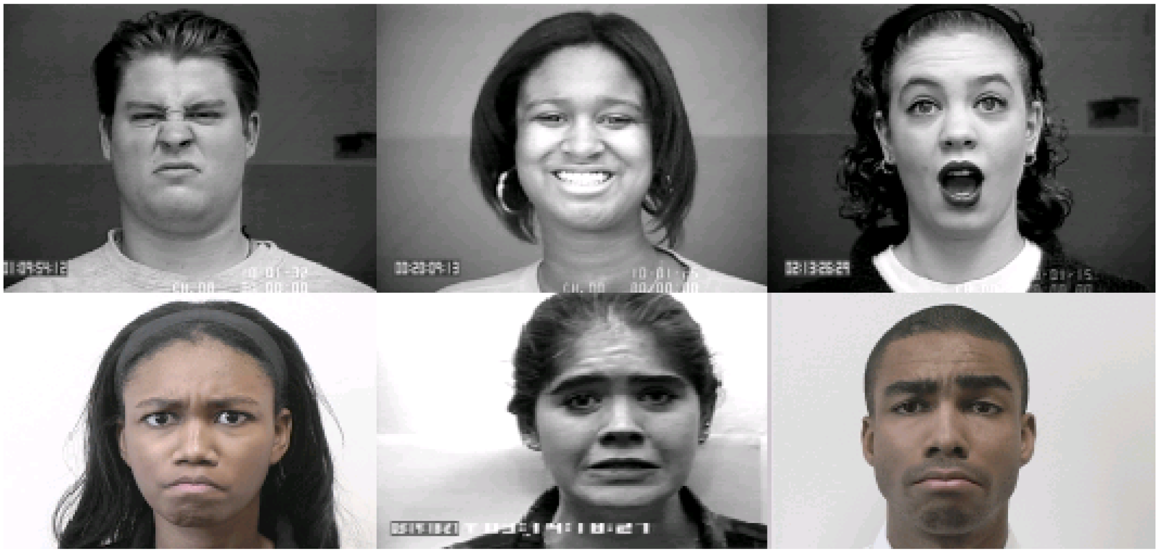


Figure 10: Example images from CK+ dataset [39].

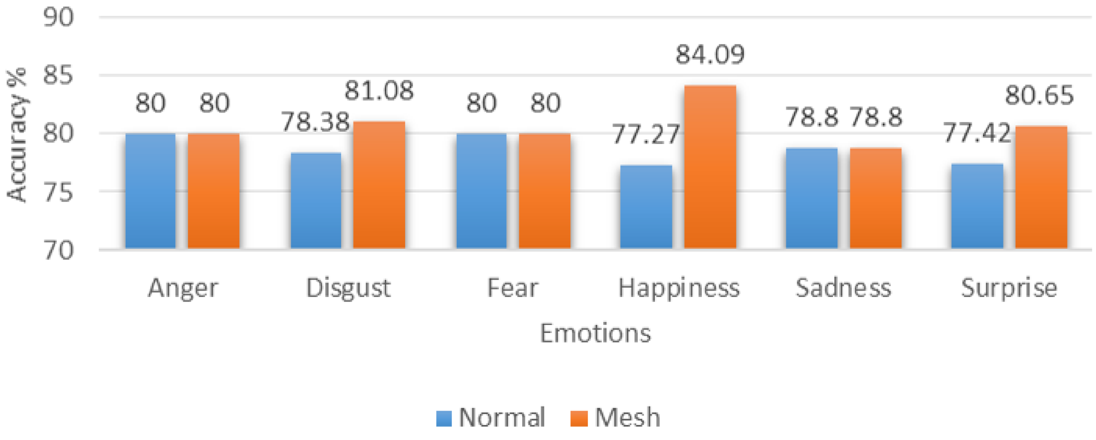


Figure 11: Accuracy comparison of each cluster for the MMI dataset, modeled by MDD mixture using both normal and mesh methods.

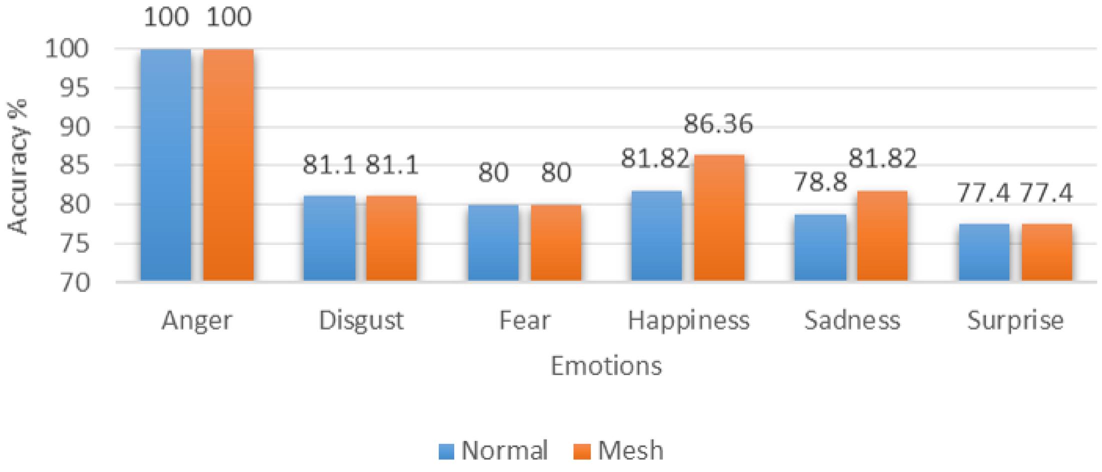


Figure 12: Accuracy comparison of each cluster for the MMI dataset, modeled by MGDD mixture using both normal and mesh methods.

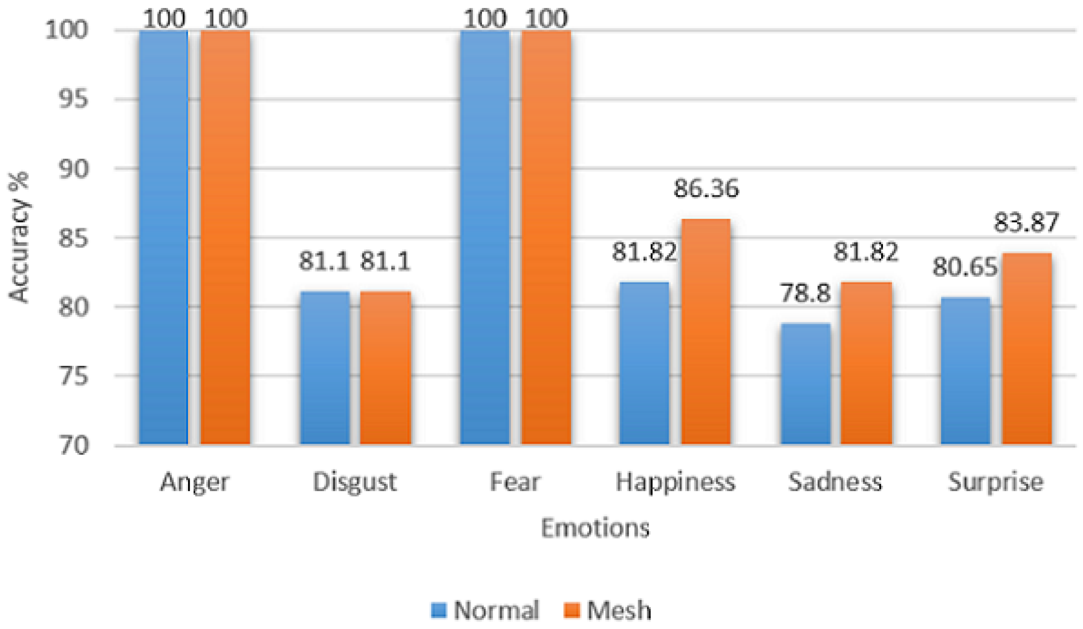


Figure 13: Accuracy comparison of each cluster for the MMI dataset, modeled by MBLD mixture using both normal and mesh methods.

Table 1: Clustering accuracy using MDD, MGDD, and MBLD with normal and mesh log-likelihood calculation for the natural scenes categorization application

Model	**MDD**		**MGDD**		**MBLD**	
Method	Normal	Mesh	Normal	Mesh	Normal	Mesh
SUN	86.02%	90.99%	88.82%	91.33%	90.68%	91.93%
Oliva & Torallba	75.46%	79.12%	78.14%	80.84%	80.07%	81.21%
Fei-Fei & Perona	72.64%	74.23%	75.67%	76.80%	76.73%	77.94%

Table 2: Facial recognition expression datasets description

Category	**MMI Dataset**		**CK+ Dataset**	
	Number of images	Portion	Number of images	Portion
Anger	150	13.16%	342	9.5%
Disgust	212	18.60%	503	12.58%
Fear	150	13.16%	417	10.43%
Happiness	255	22.37%	993	24.83%
Sadness	192	16.84%	893	22.33%
Surprise	181	15.88%	852	21%

Table 3: Clustering accuracy using MDD, MGDD, and MBLD with normal and mesh log-likelihood calculation for the facial expression recognition application

Model	**MDD**		**MGDD**		**MBLD**	
Method	Normal	Mesh	Normal	Mesh	Normal	Mesh
MMI	78.06%	81.29%	80.64%	82.58%	81.94%	83.23%
CK+	71.08%	73.96%	73.24%	74.96%	74.68%	75.40%

Chapter 4

Conclusion

In this thesis, we have introduced the usage of a novel method, for the computation of the log-likelihood function, when clustering count, multicategorial data with overdispersion, modeled by three different distributions: multinomial Dirichlet, multinomial generalized Dirichlet, and multinomial Beta-Liouville. The choice of these distributions was because of their flexibility for count data modeling as compared with the other similar distributions. In our work as a model based clustering, we used deterministic approaches such as maximum likelihood using the expectation maximization algorithm to specify the parameters of our mixture models. The strong motivation of the proposed approach was due to the numerous applications that generate such type of data. The mesh method generally reduces the error when computing the log-likelihood function, and therefore increases the clustering accuracy. The effectiveness of this technique has been shown experimentally through two applications: natural scenes categorization and facial expression recognition. In chapter two we focused on the theory of our work, discussing the different used distributions, the paired log-gamma approximation, and the mesh algorithm. In the third chapter, the performance of our model was proved experimentally via two challenging applications.

For future works, the presented procedure could be applied to other applications such as text document modeling and clustering, web mining, handwritten digit recognition, and bioinformatics including applications to metagenomics data and protein sequencing. Furthermore, more efficient optimization techniques for estimating parameters could be explored such as deterministic annealing expectation-maximization

(DAEM) approach, also using minimum description length (MDL) criterion, for determining the number of clusters.

Bibliography

[1] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[2] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[3] James H Albert and Arjun K Gupta. Bayesian estimation methods for 2×2 contingency tables using mixtures of dirichlet distributions. *Journal of the American Statistical Association*, 78(383):708–717, 1983.

[4] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[5] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[6] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

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

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

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

[10] Nizar Bouguila and Ola Amayri. A discrete mixture-based kernel for svms: Application to spam and image categorization. *Information Processing & Management*, 45(6):631–642, 2009.

[11] Nizar Bouguila and Djemel Ziou. Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization. *Journal of Visual Communication and Image Representation*, 18(4):295–309, 2007.

[12] Nizar Bouguila, Djemel Ziou, and Ernest Monga. Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, 16(2):215–225, 2006.

[13] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.

[14] Nizar Bouguila and Djerriel Ziou. Improving content based image retrieval systems using finite multinomial dirichlet mixture. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pages 23–32. IEEE, 2004.

[15] Rolf Busam and Eberhard Freitag. *Complex analysis*. Springer, 2009.

[16] Igor V Cadez, Padhraic Smyth, Geoff J McLachlan, and Christine E McLaren. Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1):7–34, 2002.

[17] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[18] George Casella and R Berger. Duxbury advanced series in statistics and decision sciences. *Statistical inference*, 2002.

[19] Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[20] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[21] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[22] Florent De Dinechin and Christoph Quirin Lauter. Optimizing polynomials for floating-point implementation. *arXiv preprint arXiv:0803.0439*, 2008.

[23] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[24] Susana Eyheramendy, David D Lewis, and David Madigan. On the naive bayes model for text categorization. 2003.

[25] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE, 2005.

[26] Paolo Giudici and Robert Castelo. Improving markov chain monte carlo model search for data mining. *Machine learning*, 50(1-2):127–158, 2003.

[27] DA Griffiths. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, pages 637–648, 1973.

[28] Rameshwar D Gupta and Donald St P Richards. Multivariate liouville distributions. *Journal of Multivariate Analysis*, 23(2):233–256, 1987.

[29] JK Haseman and LL Kupper. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, pages 281–293, 1979.

[30] Joseph M Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

[31] Peter M Hooper. Dependent dirichlet priors and optimal linear estimators for belief net parameters. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 251–259. AUAI Press, 2004.

[32] Slava M Katz. Distribution of content words and phrases in text and language modelling. *Natural language engineering*, 2(1):15–59, 1996.

[33] John D Leckenby and Shizue Kishi. The dirichlet multinomial distribution as a magazine exposure model. *Journal of Marketing Research*, 21(1):100–106, 1984.

[34] Peter Lewy. A generalized dirichlet distribution accounting for singularities of the variables. *Biometrics*, pages 1394–1409, 1996.

[35] Robert H Lochner. A generalized dirichlet distribution in bayesian life testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):103–113, 1975.

[36] Wei-Yin Loh. Symmetric multivariate and related distributions, 1992.

[37] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[38] Stephen A Lowe. The beta-binomial mixture model and its application to tdt tracking and detection. In *Proceedings of DARPA Broadcast News Workshop*, pages 127–131, 1999.

[39] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[40] David JC MacKay and Linda C Bauman Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(3):289–308, 1995.

[41] Rasmus E Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552. ACM, 2005.

[42] Diraitris Margaritis and Sebastian Thrun. A bayesian multiresolution independence test for continuous variables. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 346–353. Morgan Kaufmann Publishers Inc., 2001.

[43] G McLachlan. Peel., d. *Finite Mixture Models*, 2000.

[44] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[45] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

[46] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[47] Thomas Minka. Estimating a dirichlet distribution, 2000. *URL http://research. microsoft. com/ minka/papers/dirichlet*, 2000.

[48] James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

[49] Nagaraj K Neerchal and Jorge G Morel. An improved method for the computation of maximum likeliood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis*, 49(1):33–43, 2005.

[50] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[51] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[52] Klaas Poortema. On modelling overdispersion of counts. *Statistica Neerlandica*, 53(1):5–20, 1999.

[53] Pedro Puig and Jordi Valero. Count data distributions: some characterizations with applications. *Journal of the American Statistical Association*, 101(473):332–340, 2006.

[54] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[55] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

[56] Brian D Ripley and NL Hjort. *Pattern recognition and neural networks*. Cambridge university press, 1996.

[57] Charles H Rowe. A proof of the asymptotic series for log γ (z) and log γ (z+ a). *Annals of Mathematics*, pages 10–16, 1931.

[58] Roland T Rust and Robert P Leone. The mixed-media dirichlet multinomial distribution: A model for evaluating television-magazine advertising schedules. *Journal of Marketing Research*, 21(1):89–99, 1984.

[59] BD Sivazlian. On a multivariate extension of the gamma and beta distributions. *SIAM Journal on Applied Mathematics*, 41(2):205–209, 1981.

[60] Jaime Teevan and David R Karger. Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 18–25. ACM, 2003.

[61] H Tirri and P Kontkanen. P. myllym aki. probabilistic instance-based learning. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 507–515, 1996.

[62] Hans GC Traven. A neural network approach to statistical pattern classification by'semiparametric'estimation of probability density functions. *IEEE Transactions on Neural Networks*, 2(3):366–377, 1991.

[63] Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 737–744, 2003.

[64] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010.

[65] ET Whittaker and GN Watson. A course of modern analysis, 1990.

[66] Tzu-Tsung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

[67] Tzu-Tsung Wong. Alternative prior assumptions for improving the performance of naïve bayesian classifiers. *Data Mining and Knowledge Discovery*, 18(2):183–213, 2009.

[68] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[69] Peng Yu and Chad A Shaw. An efficient algorithm for accurate computation of the dirichlet-multinomial log-likelihood function. *Bioinformatics*, 30(11):1547–1554, 2014.

[70] Nuha Zamzami and Nizar Bouguila. Consumption behavior prediction using hierarchical bayesian frameworks. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 31–34. IEEE, 2018.