VARIATIONAL TECHNIQUES FOR MEDICAL AND IMAGE PROCESSING APPLICATIONS USING GENERALIZED GAUSSIAN DISTRIBUTION

Srikanth Amudala

A thesis
in
The Department
of
Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science Quality Systems Engineering
Concordia University
Montréal, Québec, Canada

June 2020

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Srikanth Amudala**

Entitled: **Variational techniques for medical and image processing applications using generalized Gaussian distribution**

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

Master of Applied Science Quality Systems Engineering

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

Signed by the final examining commitee:

Dr. Fuzhan Nasir _____ External examiner, BCEE

Dr. Roch Glitho _____ Internal examiner

Dr. Walter Lucia _____ Chair

Dr. Nizar Bouguila _____ Supervisor

Approved _____
Chair of Department or Graduate Program Director

_____ 2020 _____ _____

Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Variational techniques for medical and image processing applications using generalized Gaussian distribution

Srikanth Amudala

In this thesis, we propose a novel approach that can be used in modeling non-Gaussian data using the generalized Gaussian distribution (GGD). The motivation behind this work is the shape flexibility of the GGD because of which it can be applied to model different types of data having well-known marked deviation from the Gaussian shape.

We present the variational expectation-maximization algorithm to evaluate the posterior distribution and Bayes estimators of GGD mixture models. With well defined prior distributions, the lower bound of the variational objective function is constructed. We also present a variational learning framework for the infinite generalized Gaussian mixture (IGGM) to address the model selection problem; i.e., determination of the number of clusters without recourse to the classical selection criteria such that the number of mixture components increases automatically to best model available data accordingly. We incorporate feature selection to consider the features that are most appropriate in constructing an approximate model in terms of clustering accuracy. We finally integrate the Pitman-Yor process into our proposed model for an infinite extension that leads to better performance in the task of background subtraction. Experimental results show the effectiveness of the proposed algorithms.

Acknowledgments

I owe my sincere appreciation to my supervisor, Dr. Nizar Bouguila for his support and motivation in bringing the best out of me through out my graduate thesis. It will be memorable through out my carrier to work with such a cool supervisor.

I will be in-debt to all my lab mates for making this 2 year journey a memorable learning experience with their support and friendship.

Finally, I would like to thank my parents and specially my brother for supporting me with all my decisions without a slightest hesitation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical inference plays a vital role in many research areas such as computer vision, signal processing, and pattern recognition. In particular, mixture models have been widely deployed. Challenges in fitting finite mixture models include identifying the appropriate probability density function as well as the corresponding optimal number of components. Gaussian distribution has been widely used and studied with success for many applications involving computer vision, machine learning, image processing and statistical analysis [1]. However, in many real applications, Gaussian distribution fails to fit different shapes of data [2].

Recently alternative techniques have been reported in the literature to resolve the Gaussian assumption limitation. The generalized Gaussian distribution (GGD) has been proposed to provide more flexibility, by introducing a new parameter called the shape parameter. The GGD has three special cases concerning the varying shape parameter namely the Laplacian, the Gaussian, and the asymptotically uniform distributions and can be observed in Fig. 1 where β in the figure represents the shape parameter and when $\beta = 2$, the GGD becomes Gaussian.

For instance, generalized Gaussian mixture model (GGMM) has been used in [3] for buffer control, in [4, 5, 6] for texture classification and retrieval, in [7, 8, 9] for video and image segmentation, in [10] for multiresolution transmission of high-definition video, in [11] for SAR images statistics modelling, in [12] for subband decomposition of video, in [13] for denoising applications, in [14, 15] for data and image compression, in [16] for edge modeling, in [17, 18] for image thresholding, in [19, 20] to fit subband

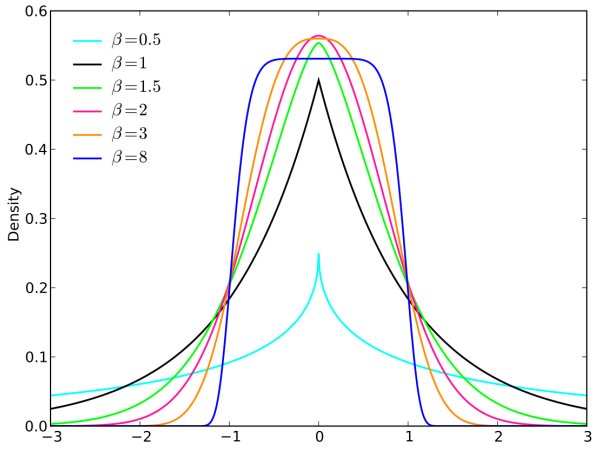


Figure 1: Generalized Gaussian distribution

histograms, in [21, 22] for speech modeling, and in [23] for multichannel audioresynthesis. The accurate modeling of wavelet coefficients distributions by GGMM was presented in [24] [25] and this property had been utilized in many signal and image processing applications which include image denoising [26], image thresholding [27], content-based image retrieval [28] and texture classification [29].

Several methods have been proposed to estimate the parameters of GGMM such as entropy matching estimation [22, 30] and maximum likelihood estimation [4, 31, 32, 33, 34] with a deterministic approach where a single distribution is considered. Maximum likelihood estimation is performed via the Expectation Maximization (EM) algorithm which has gained attention in recent times with its lower computational time. However, the EM algorithm is known for its convergence to local maxima and the tendency to overfit the model.

Solutions that incorporate Bayesian inference techniques are widely discussed in approximating intractable distributions [35]. It gives a robust hypothetical framework to utilize clustering algorithms. Markov Chain Monte Carlo (MCMC) is one of the most common techniques to estimate parameters since it is capable of accurately approximating the actual variable distribution [35] [36]. However, MCMC techniques are based on sampling to approximate the ideal distribution. This requires a large amount of computational time and resources [37]. Thus, in this thesis, we utilize

variational inference approaches [38]. Variational inference, also known as variational Bayes, is a deterministic approximation method, where, the model's posterior distribution is approximated using analytical procedures [39]. It has recently generated more interest in finite mixture models through the provision of high generalization schemes and high computation tractability. Model selection and parameter estimation can be performed simultaneously through the use of variational inference.

Model selection plays a challenging role while applying finite mixture models with a potentially inaccurate number of mixture components may result in poor generalization capability. Recent studies have tackled the problem of number of mixture components by considering a Dirichlet process (DP) prior to extend mixture models to infinity [40]. The DP permits unbounded development of the number of mixture components where it is important to fit the observations, in which the individual variables follow certain parametric distributions.

Feature selection is an important step when data are multidimensional; some features could be irrelevant and then compromise the algorithm performance as well as the clustering process. Indeed, these features do not have any discriminatory impact on the clustering. Moreover, having a high number of features increases the complexity of the model [41][42]. Thus, it is important to detect the salient features to produce efficient out comes. Consequently, in this thesis we propose a DP mixture of GGD's and employ the model proposed in [43], a feature saliency determination process, where each feature is weighted up to a probability ranging between zero and one and incorporate it into the proposed Bayesian framework.

A good alternative to DP is the Pitman-Yor process (PYP) which is a generalization to the DP prior for nonparametric Bayesian modeling. Hierarchical Bayesian nonparametric models, during the recent years, have been successfully applied in different fields such as image segmentation and language modelling [44]. The hierarchical Dirichlet process (HDP) model has shown promising results in addressing model-based clustering of grouped data with sharing clusters [45]. Using the hierarchical Pitman-Yor (HPY) process model [46], we develop a variational learning algorithm on the resulting model to estimate the parameters and apply the proposed model for background subtraction application.

1.1 Contribution

The major contributions of this thesis are as follows:

- **Variational Inference of Finite Generalized Gaussian Mixture Models:**
  We present a variational learning framework to analyze finite generalized Gaussian mixture models (GGMM). The model incorporates several mixtures that are widely used in signal and image processing applications. We present a method to evaluate the posterior distribution and Bayes estimators using the variational expectation-maximization algorithm. The effective number of components of the GGMM is determined automatically. This work has been accepted and published by Symposium Series on Computational Intelligence IEEE SSCI 2019 [47].

- **Variational Inference of Infinite Generalized Gaussian Mixture Models with Feature Selection:**
  We present a variational learning framework for the infinite generalized Gaussian mixture (IGGM) model. Infinite model addresses the model selection problem; i.e., determination of the number of clusters without recourse to the classical selection criteria such that the number of mixture components increases automatically to best model available data accordingly. We also incorporate feature selection to consider the features that are most appropriate in constructing an approximate model in terms of clustering accuracy. This work has been submitted to 2020 IEEE International Conference on Systems, Man and Cybernetics (SMC) [48].

- **Background Subtraction with a Hierarchical Pitman-Yor Process Mixture Model of Generalized Gaussian Distributions:**
  We present hierarchical Pitman-Yor process mixture of generalized Gaussian distributions for background subtraction. The Pitman-Yor process is integrated into our proposed model for an infinite extension that leads to better performance in the task of background subtraction. This work has been submitted to IEEE International Conference on Information Reuse and Integration (IRI 2020) [49].

1.2 Thesis Overview

The rest of this thesis is organized as follows:

- In chapter 2, we introduce variational inference for finite generalized Gaussian mixture models and show the results of our proposed model on real applications.
- In chapter 3, we extend our finite generalized Gaussian to the infinite case using Dirichlet process and apply feature selection for medical applications and image categorization.
- In chapter 4, we propose an infinite generalized Gaussian distribution based on the hierarchical Pitman-Yor process for background subtraction application.
- In chapter 5, we summarize our contributions.

Chapter 2

Variational Inference of Finite Generalized Gaussian Mixture Models

In this chapter, in order to tackle problems related to both Bayesian and deterministic estimation, we propose a variational approach. By considering possible distributions we assign appropriate priors to the mean and the precision of GGMM. We do not assign any prior distribution to the shape parameter of the GGMM to appropriately derive closed-form expressions.

This chapter is organized as follows. In Section 2.1, we present the variational inference of GGMM. In Section 2.2, we evaluate the performance of the proposed model on several applications.

2.1 Variational Inference of the Generalized Gaussian Mixture Model

2.1.1 Generalized Gaussian Mixture Model

The one-dimensional generalized Gaussian distribution for a vector $X \in \mathbb{R}$ with parameters μ, τ, λ is defined as follows:

$$P(X|\mu,\tau,\lambda) = \frac{\lambda\tau^{\frac{1}{\lambda}}}{2\Gamma(\frac{1}{\lambda})} e^{-\tau|(X-\mu)|^{\lambda}} \tag{1}$$

where $\tau = \left(\frac{1}{\sigma}\sqrt{\frac{\Gamma(\frac{3}{\lambda})}{\Gamma(\frac{1}{\lambda})}}\right)^{\lambda}$, $\Gamma(.)$ indicates the Gamma function given by $\Gamma(z) = \int_0^{\infty} p^{z-1}e^{-p}dp$, where z and p are real variables. The parameters μ, σ, λ denote the mean, standard deviation and the shape parameter, respectively. The parameter λ controls the shape of the probability density function. The higher the value, the flatter the probability density function indicating that λ determines the decay rate of the density function. There are two special cases, when $\lambda = 2$ and $\lambda = 1$, the GGD is reduced to the Gaussian and the Laplacian distributions, respectively. If X follows a mixture of K GGDs, then

$$P(X|\Theta) = \sum_{k=1}^{K} P(X|\mu_k, \tau_k, \lambda_k)\pi_k \tag{2}$$

where π_k ($0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$) are the mixing weights and $p(X|\mu_k, \tau_k, \lambda_k)$ is the probability density function corresponding to component k. As for the symbol $\Theta = (\epsilon, \pi)$, it refers to the entire set of parameters to be estimated where $\epsilon = (\mu_1, \tau_1, \lambda_1, ..., \mu_K, \tau_K, \lambda_K)$ and $\pi = (\pi_1, ..., \pi_K)$.

Considering N observations, $\mathcal{X} = (X_1, X_2, ..., X_N)$, and supposing that the number of components K is known, the data likelihood is denoted as follows:

$$P(\mathcal{X}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(X_n|\epsilon_k)\pi_k \tag{3}$$

where $\epsilon_k = (\mu_k, \tau_k, \lambda_k)$. For each variable X_n, let Z_n be K-dimensional vector known as the unobserved vector that assigns the appropriate mixture component that X_n belongs to. Then, Z_{nk} is equal to 0 if X_n does not belong to class k and 1, otherwise. Hence, considering $Z = (Z_1, Z_2, ..., Z_N)$ the complete-data likelihood is given by:

$$P(\mathcal{X}|\Theta, Z) = \prod_{n=1}^{N} \sum_{k=1}^{K} (P(X_n|\epsilon_k)\pi_k)^{Z_{nk}} \tag{4}$$

The EM algorithm allows to find the mixture parameters that maximize the complete data log-likelihood given by:

$$L(\mathcal{X}, Z, \Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} Z_{nk} \ln(P(X_n|\epsilon_k)\pi_k) \tag{5}$$

The assignment of X_n to the k^{th} component of the mixture can be denoted as follows [50]:

$$\hat{Z}_{nk}^{t} = \frac{P^{t-1}(X_n|\epsilon_k^{t-1})\pi_k^{t-1}}{\sum_{k=1}^{K} P^{t-1}(X_n|\epsilon_k^{t-1})\pi_k^{t-1}} \tag{6}$$

where t denotes the current step. ϵ_k^t and p_j^t are the current estimates of the parameters. A sequence of approximations to the mixture parameters Θ^t, for $t = 0, 1, ...,$ are produced by the EM algorithm until a convergence measure is fulfilled through the expectation and the maximization steps. The EM algorithm comprises of:

1. Initialize the mixture parameters.
2. E-step: Compute $\hat{Z}_{nk}^t$ (Eq. (6)).
3. M-step: Update the parameters using $\hat{\Theta}^t = \text{argmax}_{z_\Theta}\ L(\Theta, Z, \mathcal{X})$.

We note that the EM algorithm has some setbacks, like convergence to local maxima due to its dependence on initialization. A discussion on the disadvantages of the EM algorithm can be found in [51].

2.1.2 Variational Inference of the Generalized Gaussian Mixture Model

In this section, we propose a variational inference approach for the GGMM within the Variational Expectation-Maximization (VEM) framework [52] [53] to accomplish the closed-form updates and automatic determination of the number of mixture components by optimizing the Kullback–Leibler (KL) divergence between the true posterior $p(Z, \mathcal{X})$ and the approximate distribution $q(Z)$ [53]. The smaller the KL divergence, the stronger the relationship between the distributions. The KL divergence is denoted by:

$$\begin{aligned} KL(p \parallel q) &= -\int q(Z) \ln\{\frac{p(Z, \mathcal{X})}{q(Z)} - \ln p(\mathcal{X})\} dZ \\ &= -\int q(Z) \ln\{\frac{p(Z, \mathcal{X})}{q(Z)}\} dZ + \ln p(\mathcal{X}) \end{aligned} \tag{7}$$

In order to calculate the KL divergence, we need to calculate the evidence $\ln p(\mathcal{X})$. This is difficult to calculate which motivates the proposed variational inference approach. Reordering Eq. (7), we get:

$$\ln p(\mathcal{X}) = KL(p \parallel q) + \underbrace{\int q(Z) \ln\{\frac{p(Z, \mathcal{X})}{q(Z)}\} dZ}_{\text{Evidence Lower Bound}} \tag{8}$$

Maximizing the Evidence Lower Bound (ELBO) is equivalent to minimizing the KL divergence. By applying Jensen's inequality, the ELBO serves as a lower-bound for the log-evidence, $\ln p(\mathcal{X}) \geq \text{ELBO}(q)$ for any $q(Z)$, which is the approximate of the posterior. In order to maximize the ELBO, we need to choose a variational family q. The complexity of the family determines the flexibility in providing an appropriate approximation to the true posterior distribution.

We assign Normal priors for the distributions mean, and Gamma priors for the precision and shape parameters [47,48]: $\mu_k \sim N(\mu|m_0, s_0^{-1}), \tau_k \sim G(\tau|\alpha_0, \beta_0), \lambda_k \sim G(\lambda|\alpha_\lambda, \beta_\lambda)$ where $N(\mu|m_0, s_0^{-1})$ is the Normal distribution with mean m_0 and precision s_0^{-1}, $G(\tau|\alpha_0, \beta_0)$ is the Gamma distribution with shape parameter α_0 and rate parameter β_0, λ, $\mu_0, s_0, \beta_0, \alpha_0$ are the hyperparameters of the model. The posterior distributions for μ, τ, λ are defined as [50]:

$$\begin{aligned} p(\mu_k|Z,X) &\propto e^{-(\mu_k-\mu_0)^2 s_0/2 + \sum_{Z_{nk}=1} -(\tau_k|X_n-\mu_k|)^{\lambda_k}} \\ p(\tau_k|Z,X) &\propto \alpha_k^{\alpha_0-1} e^{-\beta_0\tau_k} \tau_k^{n_j} e^{\sum_{Z_{nk}=1} -(\tau_k|X_n-\mu_k|)^{\lambda_k}} \\ p(\lambda_k|Z,X) &\propto \lambda_k^{\alpha_\lambda-1} e^{-\beta_\lambda\lambda_k} \tau_k^{n_j} \left(\frac{\lambda_k}{\Gamma(1/\lambda_k)}\right)^{n_j} e^{\sum_{Z_{nk}=1} -(\tau_k|X_n-\mu_k|)^{\lambda_k}} \end{aligned} \tag{9}$$

Accordingly, we can not use the posterior distributions in their current state. To formulate the variational inference model, we denote the joint distribution of all the random variables assuming all parameters are independent as can be observed in Fig. 2:

$$p(X, Z, \pi, \mu, \tau, \lambda) = p(X|Z, \mu, \tau, \lambda)p(Z|\pi)p(\pi)p(\mu)p(\tau)p(\lambda) \tag{10}$$

For the shape parameter, a conjugate prior distribution can not be directly found. Therefore, we considered using the Taylor approximation to determine an approximate lower bound of the complete-data log-likelihood to determine whether an appropriate prior exists in the exponential family. However, the negative second-order derivative causes the function $q(\lambda)$ to be concave, resulting in an upper bound rather than a lower bound; which is required. Hence, we consider λ as a parameter and it is not assigned a prior distribution [2]. The conjugate exponential priors for μ and τ are Normal and Gamma distributions. Therefore, we specify all the priors according to:

$$\mu_k \sim N(\mu|m_k, s_k^{-1}) \tag{11}$$

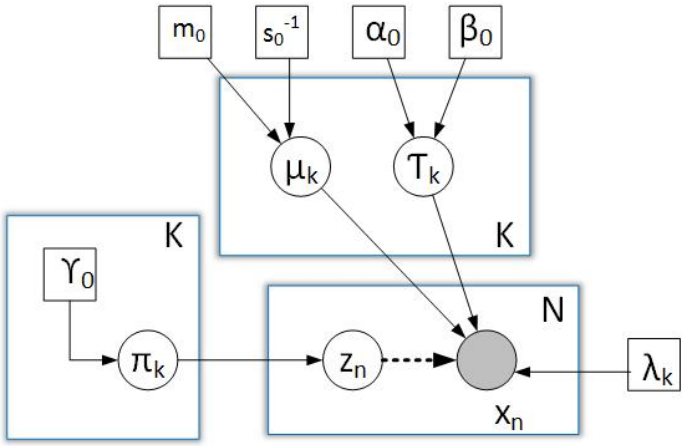


Figure 2: Graphical model for the VGGM. The filled circle, unfilled circle and square indicate observations, random variables, and parameters, respectively. The dependency among the variables is indicated by the arrows.

$$\tau_k \sim G(\tau|\alpha_k, \beta_k) \tag{12}$$

We consider the following variational distribution that factorizes into the latent variables and the parameters as:

$$q(Z, \pi, \mu, \tau, \lambda) = q(Z)q(\pi, \mu, \tau, \lambda) \tag{13}$$

$$\ln q^\star(Z) = \mathbb{E}_{\mu,\tau,\pi}[\ln p(\mathcal{X}, \pi, \mu, \tau, \lambda)] + const. \tag{14}$$

$$\ln q^\star(Z) = \mathbb{E}_{\pi}[\ln p(Z|\pi)] + \mathbb{E}_{\mu,\tau}[\ln p(\mathcal{X}|Z, \mu, \tau, \lambda)] + const. \tag{15}$$

where $\mathbb{E}$ represents the expectation with respect to the subscripted parameter and *const* denotes an additive constant. Substituting the two conditional distributions, and retaining any terms that are not dependent on Z into the constant, we have:

$$\ln q^\star(Z) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + const \tag{16}$$

where we define:

$$\begin{aligned} \ln \rho_{nk} = &\mathbb{E}_{\pi}[\ln \pi_k] + \mathbb{E}_{\mu,\tau}[\frac{1}{\lambda_k} \ln \tau_k + \ln \lambda_k - \ln 2\Gamma(1/\lambda_k) \\ &- \tau_k |X_n - \mu_k|^{\lambda_k}] \end{aligned} \tag{17}$$

Normalizing the distribution, noting for each value of n the values of Z_{nk} are binary and add up to 1 overall values of k, we obtain:

$$q^\star(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{18}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}} \tag{19}$$

The ideal solution for $q(Z)$ follows the equivalent functional form as the prior $p(Z|\pi)$. As ρ_{nk} is given by the exponential of a real quantity, the quantities ρ_{nk} will be non-negative and will sum to one. For the discrete distribution $q^\star(Z)$:

$$\mathbb{E}[z_{nk}] = r_{nk} \tag{20}$$

where r_{nk} denotes the responsibilities with the sum of all the responsibilities for the respective cluster k given by N_k as follows:

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{21}$$

Similarly, the factor in the variational posterior distribution $q(\pi, \mu, \tau, \lambda)$ is given by:

$$\ln q^\star(\pi, \mu, \tau, \lambda) = \ln q(\pi) + \sum_{k=1}^{K} q(\mu_k, \tau_k, \lambda_k) \tag{22}$$

We observe that this equation decomposes into an aggregate of terms with only π in addition to terms with μ and τ, implying that the variational posterior $q(\pi, \mu, \tau, \lambda)$ factorizes to:

$$q(\pi, \mu, \tau, \lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \tau_k, \lambda_k) \tag{23}$$

Identifying the terms that depend on π, results in:

$$\ln q^\star(\pi) = (\gamma_0 - 1) \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \pi_k + const \tag{24}$$

We recognize $q^\star(\pi)$ as a Dirichlet distribution with parameter γ:

$$q^\star(\pi) = Dir(\pi|\gamma) \tag{25}$$

where γ has components γ_k that are given by:

$$\gamma_k = \gamma_0 + N_k \tag{26}$$

$$\begin{aligned} \mathbb{E}[\ln \pi_k] &= \psi(\gamma_k) - \psi(\hat{\gamma}) \\ \hat{\gamma} &= \sum_{k=1}^{K} \gamma_k \end{aligned} \tag{27}$$

The expectation of μ with prior means m_0 and precision s_0^{-1} are denoted by:

$$\begin{aligned} \mathbb{E}[\ln q(\mu_k)] = \mathbb{E}_\tau \Bigg[\sum_{n=1}^{N} (-Z_{nk}\tau_k |X_n - \mu_k|^{\lambda_k}) - \\ \frac{s_0}{2}(\mu_k - m_0)^2 \Bigg] \end{aligned} \tag{28}$$

where $|X_n - \mu_k|^{\lambda_k}$ is expanded using the Binomial Expansion to the power 2 with the following conditions:
$if(\mu_k > X_n)$

$$|\mu_k - X_n|^{\lambda_k} = \mu_k^{\lambda_k} - \lambda_k \mu_k^{\lambda_k - 1} X_n + \frac{\lambda_k}{2}(\lambda_k - 1)\mu_k^{\lambda_k - 2} X_n^2 \tag{29}$$

$if(X_n > \mu_k)$

$$\begin{aligned} |X_n - \mu_k|^{\lambda_k} &= |X_n|^{\lambda_k} \left(1 - \frac{\mu_k}{X_n}\right)^{\lambda_k}, \\ \left(1 - \frac{\mu_k}{X_n}\right)^{\lambda_k} &= 1 - \lambda_k \frac{\mu_k}{X_n} + \frac{\lambda_k}{2}(\lambda_k - 1)\frac{\mu_k^2}{X_n^2} \end{aligned} \tag{30}$$

Substituting Eq. (29) and Eq. (30) in Eq. (28) and comparing it to the prior distribution, we obtain:

$$m_k = \frac{\frac{s_0 m_0}{2} + p_1}{s_k} \tag{31}$$

$$s_k = \frac{s_0}{2} + p_2 \tag{32}$$

where p_1, p_2 have two different cases as follows:

$$p_1 = \begin{cases} \sum_{n=1}^{N} (r_{nk}\bar{\tau}_k \frac{\lambda_k}{4}(\lambda_k - 1)\mu_k^{\lambda_k - 3} x_n^2 + \sum_{n=1}^{N} (r_{nk}\bar{\tau}_k \frac{\lambda_k}{2}\mu_k^{\lambda_k - 2} x_n)), & \text{if } X_n < m_k \\ \sum_{n=1}^{N} r_{nk}\bar{\tau}_k \lambda_k \frac{|x_n|^{\lambda_k}}{x_n}, & \text{otherwise} \end{cases}$$

$$p_2 = \begin{cases} \sum_{n=1}^{N}(r_{nk}\bar{\tau}_k\mu_k^{\lambda_k-2}), & \text{if } X_n < m_k \\ \\ \sum_{n=1}^{N}(r_{nk}\bar{\tau}_k\frac{\lambda_k}{2}(\lambda_k-1)\frac{|x_n^{\lambda_k}|}{x_n^2}), & \text{otherwise} \end{cases}$$

where $\bar{\tau}$ represents $\mathbb{E}_\tau[\tau]$. Similarly, the solution for τ is as follows:

$$\mathbb{E}[\ln q(\tau_k)] = \mathbb{E}_\mu\left[\frac{\lambda_k\tau_k^{\frac{1}{\lambda_k}}}{2\Gamma(\frac{1}{\lambda_k})}e^{-\tau_k|X-\mu_k|^{\lambda_k}} + \ln\tau_k^{\alpha_0-1} - \beta_0\tau_k\right] \tag{33}$$

$$\alpha_k = \sum_{n=1}^{N} r_{nk} + \alpha_0 - 1 \tag{34}$$

$$\beta_k = \beta_0 + \sum_{n=1}^{N} r_{nk}\,\mathbb{E}_\mu[|X_n-\mu_k|^{\lambda_k}] \tag{35}$$

$$\mathbb{E}_\mu[|X_n-\mu_k|^{\lambda_k}] = \begin{cases} |X_n|^{\lambda_k} - \lambda_k\frac{|X_n|^{\lambda_k}}{X_n}m_k + \frac{\lambda_k(\lambda_k-1)}{2}\frac{|X_n|^{\lambda_k}}{X_n^2}(\frac{1}{s_k}+m_k^2), & \text{if } X_n > \mu_k \\ \\ \mathbb{E}[|\mu_k|^{\lambda_k} - \lambda_k\mu_k^{\lambda_k-1}X_n + \frac{\lambda_k}{2}(\lambda_k-1)\mu_k^{\lambda_k-2}X_n^2], & \text{otherwise} \end{cases}$$

Then, using confluent hypergeometric function, $\mathbb{E}|\mu_k|^{\lambda_k}$ can be defined as:

$$\mathbb{E}\left[|\mu_k|^{\lambda_k}\right] = (\frac{1}{\sqrt{s_k}})^{\lambda_k}\cdot 2^{\lambda_k/2}\frac{\Gamma\left(\frac{1+\lambda_k}{2}\right)}{\sqrt{\pi}}{}_1F_1\left(-\frac{\lambda_k}{2},\frac{1}{2},-\frac{1}{2}\left(m_k\right)^2 s_k\right). \tag{36}$$

The following equation denotes the lower bound:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}[\ln P(\mathcal{X}|\Theta)] + \mathbb{E}[\ln P(Z|\pi)] + \mathbb{E}[\ln P(\pi)] \\ &+ \mathbb{E}[\ln P(\mu)] + \mathbb{E}[\ln P(\tau)] - \mathbb{E}[\ln q(Z)] \\ &- \mathbb{E}[\ln q(\pi)] - \mathbb{E}[\ln q(\mu)] - \mathbb{E}[\ln q(\tau)] \end{aligned} \tag{37}$$

The posterior distributions are obtained from the VE-step and the parameters are updated in the VM-step by augmenting the approximate lower bound $\mathcal{L}$. To approximate the parameters of the GGMM (i.e. λ), the first-order derivative of the estimated lower bound is set to zero, prompting:

$$\begin{aligned} \frac{\partial\bar{\mathcal{L}}(q,\Theta)}{\partial\lambda_k} &= \bar{\mathcal{L}}_i'(q,\Theta) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}(|X_n-\bar{\mu}_k|^{\lambda_k}\ln|X_n-\bar{\mu}_k|(\tau_k-\bar{\tau}_k) \\ &- \frac{1}{\lambda_k^2}\ln\bar{\tau}_k + \frac{1}{\lambda_k} - \frac{\Gamma'(\frac{1}{\lambda_k})}{2\Gamma(\frac{1}{\lambda_k})} + \bar{\tau}_k|X_n-\mu_k|^{\lambda_k}\ln|X_n-\mu_k|) \end{aligned} \tag{38}$$

The second-order derivative is given by:

$$\begin{aligned}\frac{\partial^2\bar{\mathcal{L}}(q,\Theta)}{\partial^2\lambda_k} &= \bar{\mathcal{L}}_i^{''}(q,\Theta) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}(2|X_n-\bar{\mu}_k|^{\lambda_k}\ln|X_n-\bar{\mu}_k|(\tau_k-\bar{\tau}_k) \\ &+ \frac{2}{\lambda_k^3}\ln\bar{\tau}_k - \frac{1}{\lambda_k^2} + \frac{1}{2}\frac{\Gamma^{'}(\frac{1}{\lambda_k})^2}{\Gamma(\frac{1}{\lambda_k})^2} - \frac{\Gamma^{''}(\frac{1}{\lambda_k})}{2\Gamma(\frac{1}{\lambda_k})} \\ &+ 2\bar{\tau}_k|X_n-\mu_k|^{\lambda_k}\ln|X_n-\mu_k|)\end{aligned} \tag{39}$$

The shape parameter is now estimated as:

$$\begin{aligned}\lambda_k^{\star} &= \lambda_k + s\Delta\lambda_k \\ \text{where } \Delta\lambda_k &= -\frac{\mathcal{L}_k^{'}(q,\Theta)}{\mathcal{L}_k^{''}(q,\Theta)}\end{aligned} \tag{40}$$

where s is determined by the backtracking line search [54]. Our complete algorithm can then be summarized as follows:

Algorithm

1. Input: $\mathcal{X}, K$, given an initial large K value.
2. Initialization: choose $\alpha_0, \beta_0, \gamma_0, m_0, s_0$ using K-means algorithm, $\lambda_k = 2$
3. Compute $\alpha_k, \beta_k, \gamma_k, m_k, s_k \leftarrow$ Initial values for each component.
4. **While** $\mathcal{L}_i - \mathcal{L}_{i-1} \leq 1e-9$
5. Compute $\ln\rho_{nk}$ using Eq. (60)
6. Generate the responsibilities r_{nk} from Eq. (61)
7. Update $\alpha_k, \beta_k, \gamma_k \leftarrow$ from Eq. (70), Eq. (71) and Eq. (26)
8. Calculate m_k, s_k from Eq. (65), Eq. (66)
9. Choose the step size s by the backtracking line search
10. Update λ_k using Eq. (96)
11. Generate lower bound $\mathcal{L}$ using Eq. (37)

12. Assign the cluster labels to the highest responsibilities in each row of the responsibility matrix.

13. **end**

2.2 Experimental results and discussion

2.2.1 Implementation details

In this section, we will be discussing about the implementation details of the proposed algorithm. The hyperparameters are set as $\alpha_0 = \mu^2/\sigma, \beta_0 = \mu/N$, given N observations. $\lambda = 2, m_0, s_0^{-1}, \gamma_0$ are initialized using K-means algorithm. Based on these initializations, we estimate the sample mean, sample precision, and shape in the i^{th} initial class. When the VEM algorithm stops, $\alpha_k, \beta_k, \gamma_k, m_k, s_k, \lambda_k$ are acknowledged as the hyperparameter and parameter estimates in the Variational GGMM (VGGMM).

2.2.2 Dataset validation

This section has two main objectives: first applying the algorithm to estimate the mixture parameters and comparing with Variational GMM (VGMM). To reach the first objective, we apply our VGGMM estimation algorithm for binary classification in medical and astrological applications involving detection of heart diseases[1] and predicting a Pulsar Star[2] and finally we apply our model in image segmentation.

Among the two data sets, the heart disease data set provides all the potential symptoms of a person having heart disease. This data set contains 76 features, however, all circulated tests allude to utilizing a subset of 14. The target field suggests the presence of heart infection within the patient. The second data set contains an example of pulsar candidates accumulated through the High Time Resolution Universe Survey. Pulsars are a phenomenal kind of Neutron star that produces radio outflow perceptible here on earth. It has picked up prominence over late occasions to mark the pulsar contender to encourage fast examination. Treating the pulsar data

[1]https://www.kaggle.com/ronitf/heart-disease-uci.

[2]https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star/downloads/predicting-a-pulsar-star.zip/1

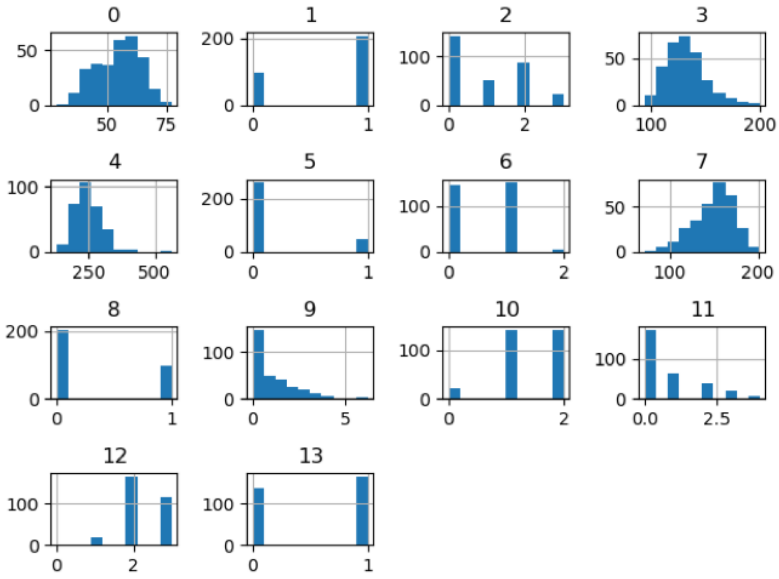


Figure 3: Histograms of Heart Disease. Histogram-0 to Histogram-12 represent the features, Histogram-13 represents the target value. X-axis indicating value range and Y-axis showing the frequency.

set as a binary classification problem makes it an ideal fit for our examination. The histograms of the input data sets are presented in Fig. 3 and Fig. 4.

We have implemented our VGGMM classifier using cross-validation with the split size of 4 for both the datasets. In order to determine the class-label of all the data points, the largest component is considered amongst the likelihood of the data points belonging to the classes. Table 1, presents the model accuracy in comparison with VGMM.

Table 1: Model accuracy comparison

Data set name	**Accuracy**		
	VGMM	***VGGMM***	***GMM***
Heart Disease UCI	41%	69.64%	52%
Predicting a Pulsar star	88%	93.2%	87%

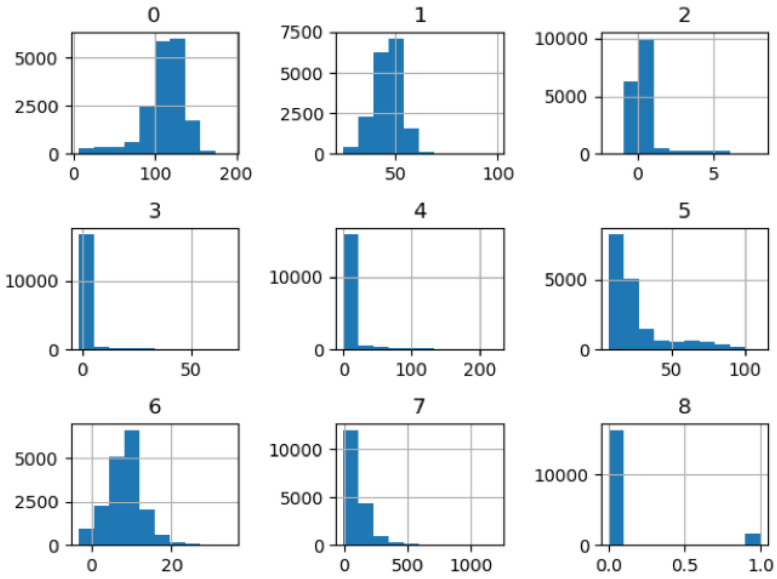


Figure 4: Histograms of Pulsar Star. Histogram-0 to Histogram-7 represent the features, Histogram-8 represents the target value. X-axis indicating value range and Y-axis showing the frequency.

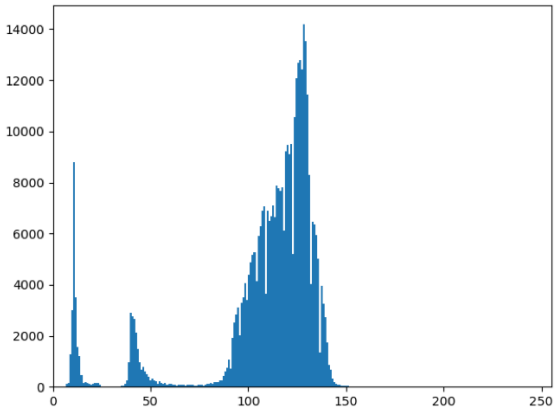
2.2.3 Image Segmentation

In computer vision, image segmentation is the process of finding the pixels with similar characteristics and clustering them to different segments. The goal of segmentation is to find similar pixels and represent the whole image in the form of segments with each segment representing pixels with similar characteristics making it easier for analysis [55][56].

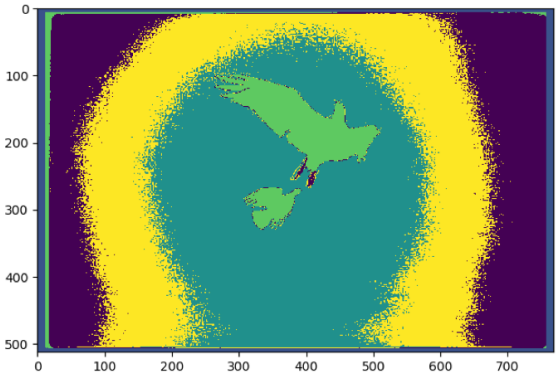
In the first experiment, we choose an image (768 x 512) with two objects in the sky to demonstrate the capability of segmenting small objects in large background (Fig. 4a). The goal is to cluster the image into two classes: the sky and the two birds. We set the number of components, $K = 5$. Comparing the outcomes for K-means algorithm, GMM, and VGMM (Fig. 4c, Fig. 4d, Fig. 4e), there is an enormous misclassification of the sky and the space between the little object and the large object. Our method, VGGMM (Fig. 4f), is able to recognize the two birds and the components effectively. Contrasted to the other methods, the wings, the tail of the little bird (red square), and the big bird are also shown in more details.

In the second experiment, we executed our estimation on a human face image (132

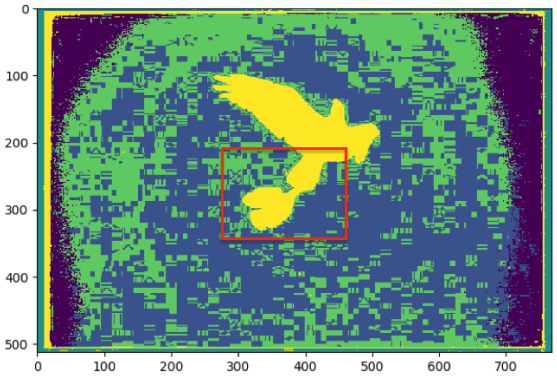
(a) Original Image



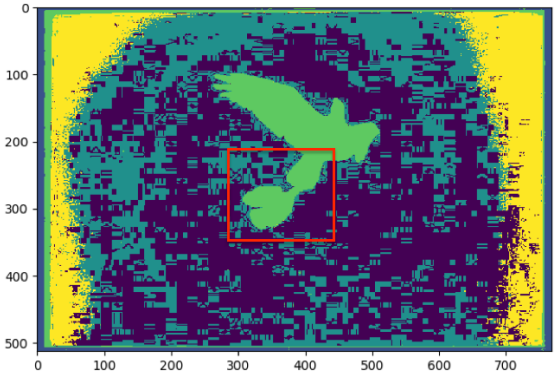
(b) histogram



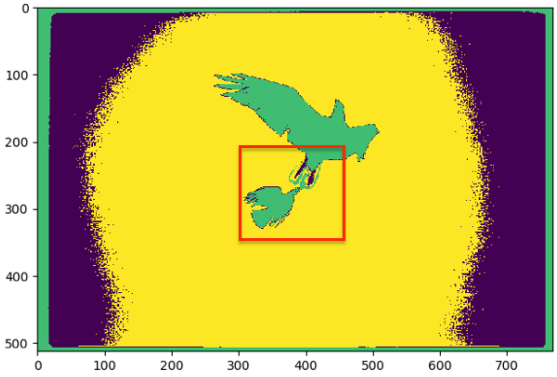
(c) K-means algorithm (K=5)



(d) GMM (K=5)



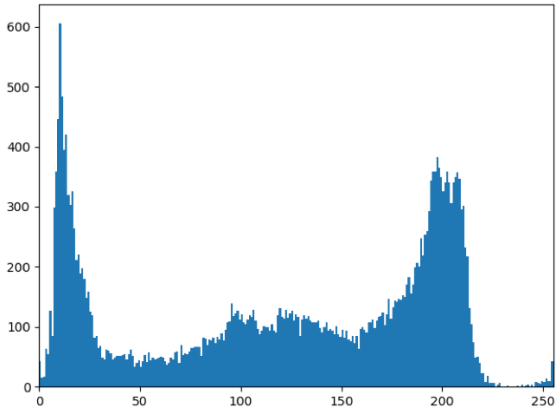
(e) VGMM (K=5)



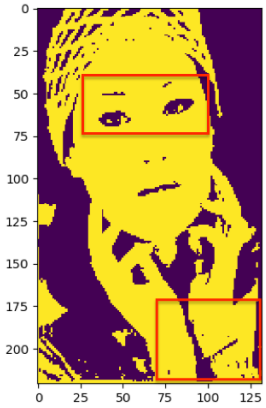
(f) VGGMM (K=5)

Figure 5: Segmentation results, Fig. 4a represents the original image.

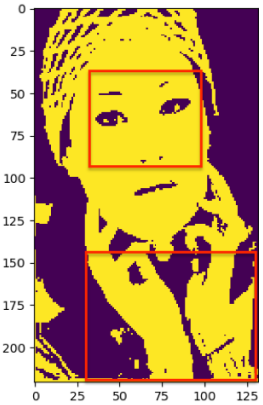
(a) Original Image



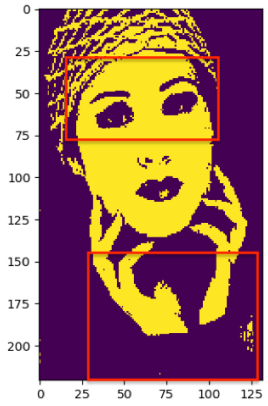
(b) histogram



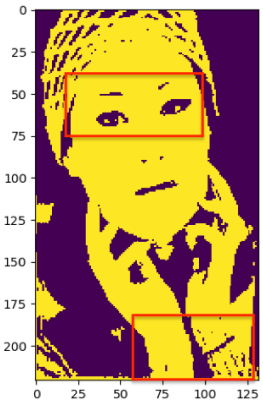
(c) K-means algorithm (K=2)



(d) GMM (K=2)



(e) VGMM (K=2)



(f) VGGMM (K=2)

Figure 6: Segmentation results, Fig. 5a represents the original image.

x 221) as shown in Fig. 5a to segment the image into two classes. In Fig. 5b, we can see the histogram of the image. We set the number of mixture components to two, $K = 2$. Comparing the result with K-means algorithm, GMM, VGMM methods, we noticed that K-means algorithm and GMM have similar results and were able to detect some features of the face. However, they contained only a part of the eyebrows and a part of the texture of clothes rather than the whole. VGMM was able to detect the eyebrows but was not able to detect the texture and the hair. Our algorithm VGGMM (Fig. 5f), was able to extract more information for image understanding.

Chapter 3

Variational Inference of Infinite Generalized Gaussian Mixture Models with Feature Selection

In this chapter, we develop a non-parametric Bayesian approach for modelling, particularly based on the Dirichlet process (DP). Here, we employ the model proposed in [43], a feature saliency determination process, where each feature is weighted up to a probability ranging between zero and one and incorporates it into the proposed Bayesian framework.

This chapter is organized as follows. In Section 3.1, we introduce the DP and stick-breaking construction. We also introduce the simultaneous clustering and feature selection algorithm and details of the proposed variational inference method. Experimental results are presented in Section 3.2.

3.1 Proposed Model

3.1.1 Dirichlet process with a stick-breaking representation

The DP is a random process with a base distribution G_0 which has probability distribution as its realization [57] and non-negative scaling parameter α. For DP construction, a random measure $G \sim DP(\alpha, G_0)$ is drawn from k-components of measure sets

$\{P_1, ..., P_k\}$ which are discrete [58]:

$$(G(P_1), ..., G(P_k)) \sim (\alpha G_0(P_1), ..., \alpha G_0(P_k)) \tag{41}$$

The learning approach is normally based on the stick-breaking process using variational inference [57]. An approximate posterior is placed on the represented set of latent variables [59]. The stick-breaking process is a representation of the DP which depends on two infinite groupings of independent and identically distributed random variables V_k and c_k, for $k \in \{1, ..., \infty\}$ [60]. Using this construction, an infinite mixture model is formed as:

$$p(V_k|\alpha) = Beta(1, \alpha) \qquad p(c_k^*|\alpha, G_0) \sim G_0 \tag{42}$$

where V_k is the stick-breaking length with concentration parameter α. c_k^* represent the atoms drawn from the base distribution G_0 independently. We define the stick-breaking representation of the random representation G as follows:

$$\pi_k = V_j \prod_{s=1}^{k-1}(1 - V_s) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{c_j^*} \tag{43}$$

δ_{c^*} is the probability concentration at c^* with weight π. The mixing weights $\pi = (\pi_k)_{k=1}^{\infty}$ are formed by breaking a unit length stick into infinite pieces with weights summing to one. Thus, the resultant has an unknown number of components that can increase as new data are observed. Thus, we have a set of observations $x = \{x_1, ..., x_N\}$ with parameters $c = \{c_1, ..., c_N\}$, where N is the total number of samples. The distribution of random measure G is formed as follows:

$$\begin{aligned} G|\{\alpha, G_0\} &\sim DP(\alpha, G_0) \\ c_n|G &\sim G \\ x_n|c_n &\sim p(x_n|c_n) \end{aligned} \tag{44}$$

where G is a random measure from a DP prior $DP(\alpha, G_0)$ and the atom c_n is independently drawn from G_0 with weight π_n given by the n^{th} stick-breaking length V_n.

We utilize the above DP mixture model with the stick-breaking process. The arbitrary variable c_n takes on c_k^* with weight π_k and the component assignment is indicated by the latent indicator variable Z_n representing the assignment of data point x_n. The generative process of the DP mixture model can be explained as follows:

- Step 1: $V_k|\alpha \sim Beta(1, \alpha), k \in \{1, ..., \infty\}$
- Step 2: $c_k^*|G_0 \sim G_0, k \in \{1, ..., \infty\}$
- Step 3: Draw the n^{th} observation, $n \in \{1, ..., N\}$
  - $Z_n|V \sim Multi(\pi)$
  - $x_n|Z_n \sim p(x_n|c_{Z_n}^*)$

From the above algorithm, the relative prevalence of the mixture is specified by the probability distribution of atoms c which is drawn from the base distribution G_0 with stick lengths V. For the observations in Step 3, the indicators Z are distributed according to a Multinomial distribution with mixing weights π generated from V.

3.1.2 Infinite generalized Gaussian mixture model

In this section we build an infinite generalized Gaussian mixture model (IGGM) utilizing the DP with the stick-breaking representation described in Section 4.1.1. In this thesis, we confine the proposed distribution to generalized Gaussian distribution (GGD) with set of parameters θ. We set a truncation level on the highest component number K of the stick-breaking representation. Given a dataset $X = \{X_1, ..., X_N\}$, if each vector $X_n = (X_{n1}, ..., X_{nD})$ is represented in a $D-$dimensional space, the truncated DP mixture model is given as follows:

$$p(X|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(X_n|\theta_k) \tag{45}$$

where $\Theta = (\pi_1, ..., \pi_K, \theta_1, ..., \theta_K)$ represents the complete set of parameters for the mixture model. $\pi = (\pi_1, ..., \pi_K)$ represents the mixing proportions which are always positive and sum up to one, and $\theta_k = (\mu_k, \tau_k, \lambda_k)$ represents the parameters of the GGD for mixture components k. The mixing weights π of the stick-breaking approach are represented as stick lengths V.

Given GGD parameters mean (μ_k), precision (τ_k) and shape (λ_k) for mixture component k, the GGD probability density function can be written as:

$$P(X_n|\theta_k) \propto \prod_{i=1}^{D} \frac{\lambda_{ik} \tau_{ik}^{\frac{1}{\lambda_{ik}}}}{2\Gamma(\frac{1}{\lambda_{ik}})} e^{-\tau_{ik}|(X_{ni} - \mu_{ik})|^{\lambda_{ik}}} \tag{46}$$

where $\tau_{ik} = \left(\frac{1}{\sigma_{ik}}\sqrt{\frac{\Gamma(\frac{3}{\lambda_{ik}})}{\Gamma(\frac{1}{\lambda_{ik}})}}\right)^{\lambda_{ik}}$, $\Gamma(.)$ denotes the gamma function given by $\Gamma(z) = \int_0^\infty p^{z-1}e^{-p}dp$ where z and p are real variables, $\mu_k = (\mu_{1k}, ..., \mu_{Dk}), \tau_k = (\tau_{1k,...,\tau_{Dk}})$, and $\lambda_k = (\lambda_{1k}, ..., \lambda_{Dk})$. The shape of the probability density function is determined by the shape parameter λ. The larger the value, the flatter the probability density function. This means that the decay rate of the density function is determined by λ. Note that for the two special cases, when $\lambda = 2$ and $\lambda = 1$, the GGD is reduced to Gaussian and Laplacian distributions, respectively. In this thesis, we assume that the covariance matrix is diagonal and each dimension of observation X_n is independent from the other dimensions.

For each variable X_n, let Z_n be a K-dimensional vector known by the unobserved vector that assigns the appropriate mixture component X_n belongs to. Then, Z_{nk} is equal to 1 if X_n belongs to class k and 0 otherwise. Hence, the complete-data likelihood is given as follows:

$$P(X|Z, \Theta) = \prod_{n=1}^{N}(p(X_n|\theta_k))^{Z_{nk}} \tag{47}$$

The mixing proportion $\pi_k = p(Z_{nk} = 1), k = \{1, ..., K\}$ indicates the probability that a data point X_n is allocated to component k. Hence, the marginal distribution over Z given a multinomial prior is given as follows:

$$p(Z|\pi) \sim Multi(\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_j^{I(Z_n=k)} \tag{48}$$

where $I(Z_n = k)$ represents the indicator function. According to Eq. (48), the mixing proportions π are represented by sticks V. Rearranging Eq. (48) gives $p(Z|V)$ as follows:

$$p(Z|V) = \prod_{n=1}^{N}\prod_{k=1}^{K}[V_k\prod_{s=1}^{k-1}(1 - V_s)]^{I(Z_n=k)} \tag{49}$$

We truncate the number of mixture components to K, with the Beta prior of stick V from Eq. (42)

$$p(V|\alpha) = \prod_{k=1}^{K} Beta(1, \alpha) = \prod_{j=1}^{K}\alpha(1 - V_k)^{\alpha-1} \tag{50}$$

3.1.3 Infinite generalized Gaussian mixture model with feature selection

Feature selection is an essential process in a mixture model as some features in the data do not necessarily contain information that is essential to clustering. We expect that each mixture component density is factorized over the features. Hence, the features are considered to be independent for each mixture component and we assume that a feature relevancy corresponds to a weight ranging between 0 and 1.

Thus, for each mixture component, we assume that a feature of X is drawn from a mixture of two univariate sub-components, as proposed in [42]. The first sub-component models relevant information since it is distinctive from all other mixture components and the second sub-component represents the "noisy" information which is common to all mixture components. Hence, we model the features with the following distribution:

$$p(X|Z,\Theta,\zeta,S) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\prod_{i=1}^{d} p(X_i|\Theta_{ik})^{s_n} p(X_i|\zeta_{ik})^{1-s_n}\right]^{z_{nk}} \quad \text{where } \Theta = \{\mu,\tau,\lambda\},\ \zeta = \{\epsilon,\delta,\Omega\} \tag{51}$$

$$p(X,Z,\pi,\mu,\tau,\lambda,\epsilon,\delta,\Omega,S) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\prod_{i=1}^{d} p(X_i|Z_{nk},\mu_{ik},\tau_{ik},\lambda_{ik})^{s_i^n} p(X_i|Z_{nk},\epsilon_{ik},\delta_{ik},\Omega_{ik})^{1-s_i^n}\right] \tag{52}$$

where ϵ, δ, and Ω are the set of parameters for the irrelevant subcomponent. The saliency of the features is expressed through the hidden variables s_i^n, where $s_i^n \in \{0,1\}$. If the value of s_i^n is one, then the i^{th} feature of X_n is generated from the relevant subcomponent; otherwise, it is generated from the irrelevant subcomponent. The distribution of the hidden variable S given the probabilities $w = \{w_i\}$ (feature saliencies) is given as follows:

$$p(S|w) = \prod_{n=1}^{N}\prod_{i=1}^{d} w_i^{s_i^n}(1-w_i)^{1-s_i^n} \tag{53}$$

3.1.4 Variational learning

In this section, we propose a variational inference framework [52] [53] for the parameters estimation of the IGGM with feature selection. Fig. 7 represents the graphical representation of our model.

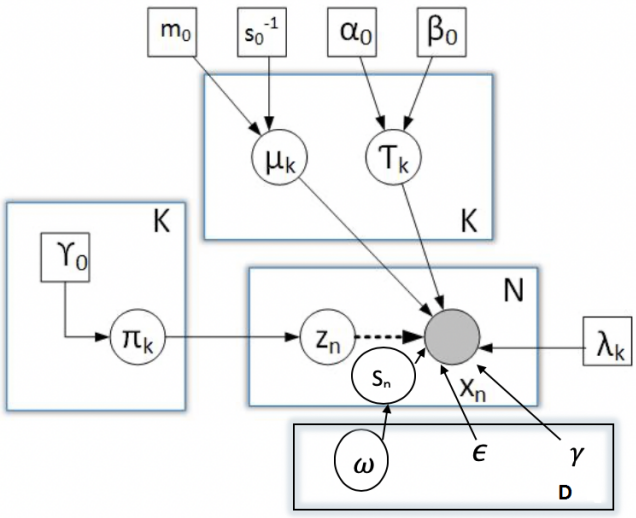


Figure 7: Graphical model for the Variational IGGM with feature selection. Filled circle, unfilled circles and squares represent observations, random variables, and parameters, respectively. The dependency among the variables is represented by directional arrows.

As discussed in the previous chapter 2 regarding the concept of variational inference, the variational distribution then factorizes into the latent variables and parameters as follows:

$$q(V, Z, \mu, \tau, \lambda, S) = \prod_{k=1}^{K} q(V_k) \prod_{n=1}^{N} q(Z_n) \prod_{k=1}^{K} \prod_{i=1}^{d} q(\mu_{ik}) q(\tau_{ik}) q(\lambda_{ik}) q(S_{in}) \tag{54}$$

where a Beta prior with parameters γ_1 and γ_2 is assigned to $q(V_k)$, $q(\mu_{ik})$ is given a normal prior with mean m_{ik} and precision s_{ik} and $q(\tau_{ik})$ is assigned a gamma prior with parameters α_{ik} and β_{ik}. $q(S_{in})$ is assigned a Bernoulli prior with parameter η_{in}. Model parameter λ_{ik} is not assigned any prior distribution [2], since the second-order

derivative of the function λ is negative making the function concave [47].

$$q^{\star}(V) = Beta(\gamma_{k1}, \gamma_{k2}) \tag{55}$$

$$q^{\star}(\mu) = N(\mu_{ik}|m_{ik}, s_{ik}^{-1}) \tag{56}$$

$$q^{\star}(\tau) = G(\tau_{ik}|\alpha_{ik}, \beta_{ik}) \tag{57}$$

$$q^{\star}(S) = \eta_{in}^{s_i^n}(1-\eta_{in})^{1-s_i^n} \tag{58}$$

Hence, the ELBO for the proposed IGGM using the mean field assumption is given as follows:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N}(\mathbb{E}[\ln p(X_n|\Theta)] + \mathbb{E}[\ln p(Z_n)]) \\ &+ \mathbb{E}[\ln p(\mu)] + \mathbb{E}[\ln p(\tau)] + \mathbb{E}[\ln p(S)] + \mathbb{E}[\ln p(V)] \\ &- \mathbb{E}[\ln q(V, Z, \mu, \tau, \lambda, S)]\end{aligned} \tag{59}$$

By applying Eq. (54) for every factor, the optimal solution of the variational posterior for all the factors is given as follows:

$$\begin{aligned}\ln \rho_{nk} = &\mathbb{E}_V[\ln V_k] + \sum_{m=1}^{k-1}\mathbb{E}_V[\ln(1-V_m)] + \mathbb{E}_{\mu,\tau,s}\left[s_{nk}\left(\ln \frac{\lambda_k \tau_k^{\frac{1}{\lambda_k}}}{2\Gamma(\frac{1}{\lambda_k})} - \tau_k|X-\mu_k|^{\lambda_k}\right) + \right. \\ &\left.(1-s_{nk})\left(\ln \frac{{}_k\Lambda_k^{\frac{1}{\delta_k}}}{2\Gamma(\frac{1}{k})} - \Lambda_k|X_n-\delta_k|^{k}\right)\right]\end{aligned} \tag{60}$$

The variational parameters r_{nk}, γ_1, γ_2, m_{ik}, s_{ik}^{-1},α_{ik}, β_{ik} and η_{in} are obtained by maximizing and determining the densities involved in q. The variational parameters are defined using the expected values of $z_{nk}, \mu_{ik}, \tau_{ik}, s_i^n, V_k$ and corresponding functions of these parameters. The following equations are obtained after deriving the expectation from $q^{\star}(V), q^{\star}(\mu), q^{\star}(\tau)$ and $q^{\star}(S)$ as follows:

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}} \tag{61}$$

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{62}$$

$$\gamma_{k1} = 1 + \sum_{n=1}^{N} r_{nk} \tag{63}$$

$$\gamma_{k2} = \alpha + \sum_{n=1}^{N} \sum_{m=k+1}^{K} r_{nm} \tag{64}$$

$$m_{ik} = \frac{\frac{s_0 m_0}{2} + t_1}{s_{ik}} \tag{65}$$

$$s_{ik} = \frac{s_0}{2} + t_2 \tag{66}$$

$$\eta_{in} = \frac{w_i \hat{\eta_{in}}}{w_i \hat{\eta_{in}} + (1 - w_i)\varepsilon_{in}} \tag{67}$$

$$\begin{aligned} \hat{\eta}_{in} = \exp \Bigg\{ & \frac{1}{2} \sum_{k=1}^{K} r_{nk} [\psi(\alpha_{ik}) - \log \beta_{ik}] \\ & - \frac{1}{2} \sum_{k=1}^{K} r_{nk} \frac{\alpha_{ki}}{\beta_{ki}} [(x_i^n - m_{ik})^2 + \tau_{ik}] \Bigg\} \end{aligned} \tag{68}$$

$$\varepsilon_{in} = \exp \left\{ -\frac{1}{2} \gamma_i (x_i^n - \epsilon_i)^2 + \frac{1}{2} \log \gamma_i \right\} \tag{69}$$

where t_1, t_2 have two different cases as follows:

$$t_1 = \begin{cases} \sum_{n=1}^{N} (r_{nk} \bar{s}_n \bar{\tau}_{ik} \frac{\lambda_{ik}}{4} (\lambda_{ik} - 1) \mu_{ik}^{\lambda_{ik} - 3} x_n^2 + \\ \sum_{n=1}^{N} (r_{nk} \bar{s}_n \bar{\tau}_{ik} \frac{\lambda_k}{2} \mu_{ik}^{\lambda_k - 2} x_n)), \text{if } X_n < m_k \\ \\ \sum_{n=1}^{N} r_{nk} \bar{s}_n \bar{\tau}_k \lambda_k \frac{|x_n|^{\lambda_k}}{x_n}, \text{otherwise} \end{cases}$$

$$t_2 = \begin{cases} \sum_{n=1}^{N} (r_{nk} \bar{s}_n \bar{\tau}_{ik} \mu_{ik}^{\lambda_{ik} - 2}), \text{if } X_n < m_{ik} \\ \\ \sum_{n=1}^{N} (r_{nk} \bar{s}_n \bar{\tau}_{ik} \frac{\lambda_{ik}}{2} (\lambda_{ik} - 1) \frac{|x_n^{\lambda_{ik}}|}{x_n^2}), \text{otherwise} \end{cases}$$

where $\bar{\tau}$ represents $\mathbb{E}_\tau[\tau]$.

$$\alpha_{ik} = \sum_{n=1}^{N} \bar{s}_n r_{nk} + \alpha_0 - 1 \tag{70}$$

$$\beta_{ik} = \beta_0 + \sum_{n=1}^{N} \bar{s}_n r_{nk}\, \mathbb{E}_\mu[|X_n - \mu_{ik}|^{\lambda_{ik}}] \tag{71}$$

$$\mathbb{E}_\mu[|X_n - \mu_{ik}|^{\lambda_{ik}}] = \begin{cases} |X_n|^{\lambda_{ik}} - \lambda_{ik}\frac{|X_n|^{\lambda_{ik}}}{X_n} m_{ik} + \\ \frac{\lambda_{ik}(\lambda_{ik}-1)}{2}\frac{|X_n|^{\lambda_{ik}}}{X_n^2}(\frac{1}{s_{ik}} + m_{ik}^2), \\ \text{if } X_n > \mu_{ik} \\ \\ \mathbb{E}[|\mu_{ik}|^{\lambda_{ik}} - \lambda_{ik}\mu_v^{\lambda_{ik}-1} X_n + \\ \frac{\lambda_{ik}}{2}(\lambda_{ik}-1)\mu_{ik}^{\lambda_{ik}-2} X_n^2], \text{otherwise} \end{cases}$$

Then using the confluent hypergeometric function results in:

$$\begin{aligned} \mathbb{E}\left[|\mu_{ik}|^{\lambda_{ik}}\right] = \\ & (\frac{1}{\sqrt{s_{ik}}})^{\lambda_{ik}} \cdot 2^{\lambda_{ik}/2} \frac{\Gamma\left(\frac{1+\lambda_{ik}}{2}\right)}{\sqrt{\pi}} {}_1 \\ & F_1\left(-\frac{\lambda_{ik}}{2}, \frac{1}{2}, -\frac{1}{2}\left(m_{ik}\right)^2 s_{ik}\right). \end{aligned} \tag{72}$$

$$\begin{aligned} \mathbb{E}[\ln V_k] &= \psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2}) \\ \mathbb{E}[\ln(1 - V_k)] &= \psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2}) \end{aligned} \tag{73}$$

After the maximization of lowerbound $\mathcal{L}$ with respect to Q, the second step of the method requires maximization of $\mathcal{L}$ with respect to w_i, ϵ_i, and γ_i. Setting the derivative of $\mathcal{L}$ with respect to the parameters equal to zero results in the following update rules:

$$w_i = \frac{1}{N}\sum_{n=1}^{N} \eta_{in} \tag{74}$$

$$\epsilon_i = \frac{\sum_{n=1}^{N} \eta_{in} x_i^n}{\sum_{n=1}^{N} \eta_{in}} \tag{75}$$

$$\frac{1}{\gamma_i} = \frac{\sum_{n=1}^{N} \eta_{in}(x_i^n - \epsilon_i)^2}{\sum_{n=1}^{N} \eta_{in}} \tag{76}$$

Given the posterior distributions from the variational expectation (E)-step, the variational maximization (M)- step updates the parameters by maximizing the approximate lower bound $\mathcal{L}$. To estimate the parameters of the GGD, i.e. λ,

$$\begin{aligned} \lambda_{ik}^{\star} &= \lambda_{ik} + \iota \Delta\lambda_{ik} \\ \text{where } \Delta\lambda_{ik} &= -\frac{\mathcal{L}_{ik}^{'}(q,\Theta)}{\mathcal{L}_{ik}^{''}(q,\Theta)} \end{aligned} \tag{77}$$

where ι is determined by the backtracking line search [54].

Algorithm 1 Variational learning of infinite generalized Gaussian mixture model with feature selection

1. **Initialization:** Initialize the truncation level K and hyperparameters α_{i0}, β_{i0}, m_{i0}, s_{i0} and r_{nk} using K-means algorithm, $\lambda_{ik} = 2$.
2. Initialize, s_i^n, w_i, ϵ_i, γ_i and η_{in} and compute α_{ik}, β_{ik}, m_{ik} and s_{ik}.
3. **loop**
   - i Update the irrelevant assignments w_i, ϵ_i, γ_i, and η_{in} from the posteriors using Eq. (74), Eq. (75) Eq. (76), Eq. (67) and Eq. (96).
   - ii Calculate m_{ik} and s_{ik} from Eq. (65) and Eq. (66).
   - iii Choose the step size ι by the backtracking line search and update λ_{ik} using Eq. (96).
   - iv The convergence criteria is reached when the difference of the current value of joint posteriors and the previous value is less than $1e^{-9}$. Otherwise, repeat above loop until convergence.

   end
4. Compute the expected value of stick length V_j and the value of mixing proportions using Eq. (43).
5. Detect the ideal number of mixture components K by eliminating the components with small mixing coefficients close to zero.

3.2 Experimental results and discussion

In this section, we evaluate the proposed variational IGGM model using image categorization and a medical application. We compare the effectiveness of the model

based on Gaussian mixture model (GMM) and variational Gaussian mixture model (VGMM). For efficient computation, we set $\Omega = 2$ for the irrelevant subcomponent to be a Gaussian distribution.

3.2.1 Image categorization

Image categorization plays an important role in automation and multimedia applications where identifying patterns is vital [61]. In our experimental setup, we choose the Caltech 101 objects dataset [62]. Among the 101 categories, we choose four categories: Bikes, Yin Yang, Sunflowers and Aeroplanes. All the categories have 60 images each to have a balanced dataset. Sample images are shown in Fig. 8. Also, to evaluate the robustness of our model, all the categories that are considered have a similar landscape.

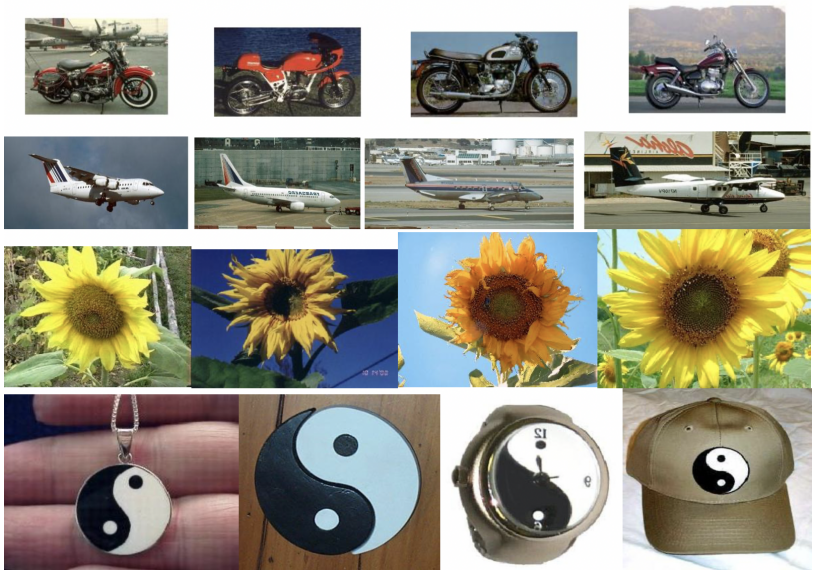


Figure 8: Caltech 101 categories utilized in this chapter (top to bottom rows): Motorbike, Aeroplane, Sunflower, Yin Yang.

To implement our model, we initially extract features and utilize the bag of visual words (BoVW) representation [63][64]. Some of the most commonly utilized descriptors are Scale-Invariant Feature Transform (SIFT) [65], Speeded Up Robust Features (SURF) [66], Histogram of Oriented Gradients (HOG) [67]. In this chapter, we use SIFT features for representations of the Caltech 101 dataset. SIFT feature extraction

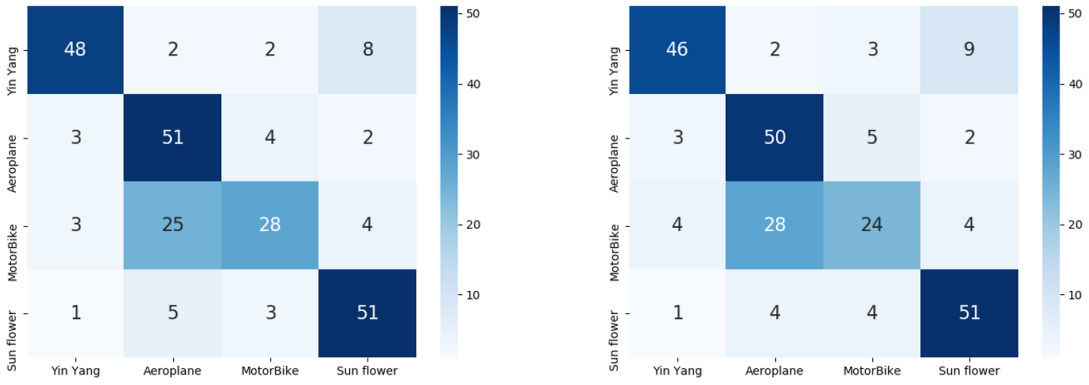
has the target of decreasing the subsequent computational complication and facilitating credible and accurate recognition for unknown new data. Our BoVW approach consists of 200 features. Consequently, we first extract the features from the images and perform K-means clustering over the extracted SIFT descriptors to form the bag of the words feature vector for each image.

Our experiments comprise of clustering with no training stage as information is infused into the algorithm with no prior knowledge about the observation labels. As outlined in Fig. 8, the Caltech 101 dataset for a given label has many number of images with different objects along with the focused object. We initialize the input dataset using K-means algorithm and start with one mixture component ($K = 1$). The proposed algorithm, denoted in Algorithm 1, then iterates until convergence. We evaluate the effectiveness of the model in terms of the accuracy, recall and the precision metrics which are defined as accuracy = (TP + TN)/Total no of observations, recall = TP/(TP +FN) and precision = TP/(TP + FP) where TP, TN, FP, and FN represent the total number of true positives, true negatives, false positives, and false negatives respectively.

Fig. 9 depicts the confusion matrix of the variational IGGM with and without feature selection. Our results show that the model has misclassified Aeroplane as MotorBike because of the high similarity of the landscape. Nonetheless, Table 2 shows that our model outperforms the other comparing models as well as the variational IGGM without feature selection. We can observe that VGMM resulted in a much lower accuracy and precision than any other model due to overfitting. Incorporating feature selection into the IGGM has improved the accuracy by 3%.

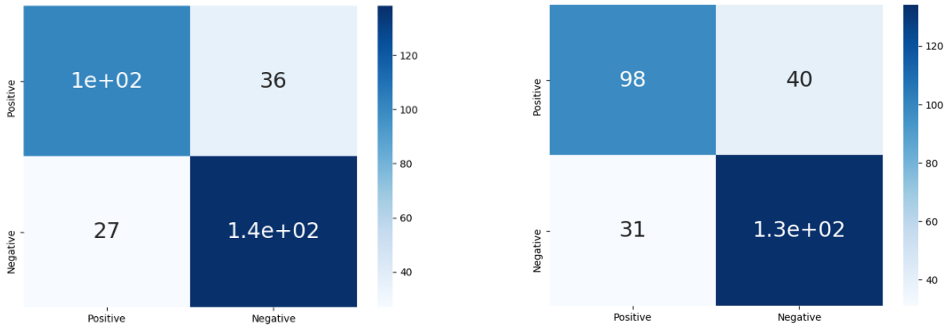
Table 2: Results for image categorization application with the Caltech 101 dataset and 200 features.

Method	**Precision**(%)	**Recall**(%)	**Accuracy**(%)
GMM	33.31	38.43	38.34
VGMM	14.10	25.61	25.41
IGGM without feature selection	72.51	71.10	71.40
IGGM with feature selection	75.12	74.67	74.51



(a) with feature selection. (b) without feature selection.

Figure 9: Confusion matrices of variational IGGM model for for Caltech 101 dataset.



(a) with feature selection. (b) without feature selection.

Figure 10: Confusion matrices of variational IGGM model for heart disease dataset.

3.2.2 Heart Disease Detection

For the second application, we apply our proposed variational IGGM estimation algorithm with feature selection in medical applications involving detection of heart diseases. The heart disease data set provides all the potential symptoms of a person with positive heart disease.

We have implemented our variational IGGM model with and without feature selection starting with $K = 1$. The label for each data point is determined with the largest component among the likelihood of the data point belonging to the classes.

Fig. 12 represents the confusion matrix results of the variational IGGM model with and without feature selection. We can observe that inclusion of feature selection increased true positives significantly by reducing the false positives when compared

Table 3: Results of Heart Disease UCI dataset.

Method	**Precision**(%)	**Recall**(%)	**Accuracy**(%)
GMM	50.43	58.31	51.22
VGMM	57.07	62.31	59.10
IGGM without feature selection	77.10	77.10	76.10
IGGM with feature selection	79.33	79.34	79.33

with the model without feature selection which is crucial in any medical-related application.

Table 3 presents the precision, recall and model accuracy of the three algorithms. Although VGMM performed better than GMM due to relatively less number of features, we can see that the variational IGGM model performed better than all the other models and the inclusion of feature selection resulted in an improvement of 3% in precision, recall and accuracy.

Chapter 4

Background Subtraction with a Hierarchical Pitman-Yor Process Mixture Model of Generalized Gaussian Distributions

Gaussian mixture models (GMM) are widely used for video background subtraction [1]; however, the foreground and the background pixels are not necessarily always distributed as a Gaussian [68]. In this work, we take advantage of the flexibility of the generalized Gaussian distribution (GGD) to fit the foreground and the background pixels [47].

In this chapter, we use the hierarchical Pitman-Yor (HPY) process model [46], we develop a variational learning algorithm on the resulting model to estimate the parameters and apply the proposed model for background subtraction. The rest of the chapter is organized as follows. In the next section, we present HPY process mixture model with GGD. The model learning is presented in Section 4.2. Section 4.3 is devoted to the experimental results.

4.1 Model specification

4.1.1 Hierarchical Pitman-Yor process mixture model

The PYP for a random distribution G with a base distribution H is defined with two parameters; namely, a discount parameter ι_a and a concentration parameter ι_b, satisfying $0 < \iota_a < 1, \iota_b > -\iota_a$ and given by [69]:

$$G \sim PYP(\iota_a, \iota_b, H) \tag{78}$$

$\iota_a = 0$ is a special case of DP with concentration parameter ι_b. The HPY process is an extension to the PYP with a Bayesian hierarchy and the base measure is itself distributed according to a PYP prior. The HPY process consists of a base distribution G_0 and a group-level distribution G_j which are formed using the stick-breaking construction. It gives an explicit representation of the HPY which depends on two infinite random variables $\Phi'_k = \{\Phi'_1, ..., \Phi'_\infty\}$ and $\kappa_k = \{\kappa_1, ..., \kappa_\infty\}$ which are independent and are distributed identically. The stick-breaking construction of the base distribution G_0 can be described as follows [69]:

$$\begin{aligned} &\kappa_k \sim H, \quad \Phi'_k \sim \text{Beta}\left(1 - \iota_a, \iota_b + k\iota_a\right) \\ &\Phi_k = \Phi'_k \prod_{l=1}^{k-1} \left(1 - \Phi'_l\right), \quad G_0 = \sum_{k=1}^{\infty} \Phi_k \delta_{\kappa_k} \end{aligned} \tag{79}$$

where κ_k is the set of independent random samples distributed according to the base distribution H. Φ_k represents the stick-breaking weights, $\sum_{k=1}^{\infty} \Phi_k = 1$ and δ_{κ_k} is an atom at κ_k. The stick lengths Φ' are defined using the two parameters ι_a and ι_b of the Beta distribution. The stick-breaking representation of the group-level PYP process is defined as follows:

$$\begin{aligned} &\psi_{jt} \sim G_0, \quad p'_{jt} \sim \text{Beta}\left(1 - \mathfrak{B}_a, \mathfrak{B}_b + t\mathfrak{B}_a\right) \\ &p_{jt} = p'_{jt} \prod_{s=1}^{t-1} \left(1 - p'_{js}\right), \quad G_j = \sum_{t=1}^{\infty} p_{jt} \delta_{\psi_{jt}} \end{aligned} \tag{80}$$

where p_{jt} represents the stick-breaking weights and satisfies $\sum_{t=1}^{\infty} p_{jt} = 1$. p'_{jt} is the stick-breaking lengths used to recursively cut a unit length stick into infinite number of pieces. The stick lengths p'_{jt} follow a Beta prior and are defined using two parameters $\mathfrak{B}_a$ and $\mathfrak{B}_b$. ψ_{jt} is distributed according to the base distribution G_0 and $\delta_{\psi_{jt}}$ represents the corresponding realization concentrated at ψ_{jt}.

We assign global-level indicator variables I such that $I_{jtk} \in \{0,1\}$. For each ψ_{jt}, $I_{jtk} = 1$ if ψ_{jt} maps to the base-level atom κ_k which is indexed by k; $I_{jtk} = 0$, otherwise. Hence, we can represent $\psi_{jt} = \kappa_k^{W_{jtk}}$. The indicator variable follows a Multinomial distribution with stick parameter Φ and is defined as follows:

$$p(I|\Phi) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \text{Multi}(\Phi) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \Phi_k^{I_{jtk}} \tag{81}$$

As Φ is a function of Φ' according to the stick-breaking construction in Eq. (79), we can rewrite Eq. (81) as follows:

$$p\left(I|\Phi'\right) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\Phi'_k \prod_{l=1}^{k-1} \left(1 - \Phi'_l\right) \right]^{I_{jtk}} \tag{82}$$

The prior for Φ' is drawn from a Beta distribution described in Eq. (79) and can be given as follows:

$$p\left(\vec{\Phi}'\right) = \prod_{k=1}^{\infty} \frac{\Gamma\left(1 - \iota_{ak} + \iota_{bk} + k\iota_{ak}\right)}{\Gamma\left(1 - \iota_{ak}\right)\Gamma\left(\iota_{bk} + k\iota_{ak}\right)} \left(1 - \Phi'_k\right)^{\iota_{bk} + k\iota_{ak} - 1} \Phi_k'^{-\iota_{ak}} \tag{83}$$

We construct the HPY process mixture as a factor associated with the observation X_{ji}, where i indexes the observations within each j^{th} group of the grouped dataset. The HPY process mixture generates θ_{ji} as a factor to every observation of X_{ji}, and $\theta_j = (\theta_{j1}, \theta_{j2}, ...)$ and are distributed according to G_j of the PYP. Hence, the likelihood function is given as follows:

$$\theta_{ji} \left| G_j \sim G_j, \quad X_{ji} \right| \theta_{ji} \sim F\left(\theta_{ji}\right) \tag{84}$$

where $F(\theta_{ji})$ represent the distribution of X_{ji} given the factor θ_{ji}. The base distribution H of G_0 gives the prior for θ_{ji}. As per this setup, each group j is related with a mixture model, and as the atoms κ_k are shared among all G_j; therefore, the mixture components are also shared among the mixture models. As each factor θ_{ji} is distributed according to G_j with values ψ_{jt} and probability p_{jt}. We introduce one more latent indicator variable W following the Multinomial distribution as:

$$p(W|p) = \prod_{j}^{M} \prod_{i}^{N} \prod_{t}^{\infty} p_{jt}^{W_{jit}} \tag{85}$$

Hence, for each θ_{ji}, we place an indicator variable $W_{jit} \in \{0,1\}$ where $W_{jit} = 1$ if θ_{ji} belongs to component t and maps to the group-level atom ψ_{jt}; otherwise, $W_{jit} = 0$.

Therefore, we have $\theta_{ji} = \psi_{jt}^{W_{jit}}$. Since ψ_{jt} maps to the global-level atom κ_k, we can also write $\theta_{ji} = \psi_{jt}^{W_{jit}} = \kappa_k^{W_{jtk} I_{jtk}}$.

According to the stick-breaking construction in Eq. (80), rewriting Eq. (85) results in:

$$p(W|p') = \prod_{j}^{M} \prod_{i}^{N} \prod_{t}^{\infty} [p'_{jt} \prod_{s=1}^{t-1} (1 - p'_{js})]^{W_{jit}} \tag{86}$$

The prior for p' is given by a Beta distribution described in Eq. (80) and can be given as follows:

$$\begin{aligned} p\left(\vec{p'}\right) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} & \frac{\Gamma\left(1 - \mathfrak{B}_{ajt} + \mathfrak{B}_{bjt} + t\mathfrak{B}_{ajt}\right)}{\Gamma\left(1 - \mathfrak{B}_{ajt}\right) \Gamma\left(\mathfrak{B}_{bjt} + k\mathfrak{B}_{ajt}\right)} \\ & \left(1 - p'_{jt}\right)^{\mathfrak{B}_{bjt} + t\mathfrak{B}_{ajt} - 1} p_{jt}^{\prime - \mathfrak{B}_{ajt}} \end{aligned} \tag{87}$$

4.1.2 HPY mixture of generalized Gaussian distributions

In this thesis, we restrict the base distribution H in Eq. (78) to GGD. Given the dataset X having N random vectors divided into M groups, where each D dimensional observation $X_{ji} = (X_{ji1}, ..., X_{jiD})$ is drawn from a HPY process mixture model of GGD's with parameters $\mu_k = (\mu_{1k}, ..., \mu_{Dk})$, $\tau_k = (\tau_{1k,...,\tau_{Dk}})$, and $\lambda_k = (\lambda_{1k}, ..., \lambda_{Dk})$. Thus, the likelihood function with the latent indicators can be given as follows [69]:

$$\begin{aligned} p(X|W, I, \mu, \tau, \lambda) &= \prod_{j=1}^{M} \prod_{i=1}^{N} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} p\left(X_{ji} | \mu_k, \tau_k, \lambda_k\right)^{W_{jit} I_{jtk}} \\ &= \prod_{j=1}^{M} \prod_{i=1}^{N} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\prod_{d=1}^{D} \frac{\lambda_{kd} \tau_{kd}^{\frac{1}{\lambda_{kd}}}}{2\Gamma(\frac{1}{\lambda_{kd}})} e^{-\tau_{kd} |(X_{jid} - \mu_{kd})|^{\lambda_{kd}}} \right]^{W_{jit} I_{jtk}} \end{aligned} \tag{88}$$

$\Gamma(.)$ denotes the gamma function given by $\Gamma(z) = \int_0^\infty p^{z-1} e^{-p} dp$, where z and p are real variables. Normal $\mathcal{N}$ and Gamma $\mathcal{G}$ priors are assigned to the parameters μ and τ with hyperparameters p, q, m, and s respectively as follows:

$$\begin{aligned} \mu_{kd} &\sim \mathcal{N}\left(p_{kd}, q_{kd}^{-1}\right) \\ \tau_{kd} &\sim \mathcal{G}\left(m_{kd}, s_{kd}\right) \end{aligned} \tag{89}$$

No prior distribution is considered for shape (λ) parameter [47].

4.2 Variational inference

In this section, we use the already presented variational inference from the previous chapter 2 to approximate a distribution $q(\Theta)$ for the true posterior $p(\Theta|X)$, where $\Theta = \{I, \Phi', W, p', \mu, \tau, \lambda\}$ indicates the set of latent variables in the HPY process GGM (HPYPGGM). Thus, the mean field variational inference of HPYPGGM is given by:

$$q\left(I, \Phi', W, p', \mu, \tau, \lambda\right) = q(I)q\left(\Phi'\right)q(W)q\left(p'\right)q(\mu)q\left(\tau\right)q\left(\lambda\right) \tag{90}$$

In our algorithm, we truncate the variational approximation of the base distribution G_0 at $K : \beta'_K = 1, \beta_k = 0$ when $k > K$, satisfying the condition $\sum_{k=1}^{K} \beta_k = 1$. Similarly for the variational approximate G_j at $T : p'_{jT} = 1$, $p_{jt} = 0$ when $t > T$ and, $\sum_{t=1}^{T} p_{jt} = 1$. The variational parameters K and T are optimized during the variational learning process. Next, considering the suitable family of variational approximations, we can have the distributions for the parameters as follows:

$$\begin{aligned}
q(I) &= \prod_{j}^{M}\prod_{t}^{T}\prod_{k}^{K} \text{Multi}\left(I_{jtk}|\varphi_{jtk}\right) \\
q(W) &= \prod_{j}^{M}\prod_{i}^{N}\prod_{t}^{T} \text{Multi}\left(W_{jit}|\varrho_{jit}\right) \\
q\left(\Phi'\right) &= \prod_{k}^{K} \text{Beta}\left(\Phi'_k|c_k, d_k\right) \\
q\left(p'\right) &= \prod_{j}^{M}\prod_{t}^{T} \text{Beta}\left(p'_{jt}|e_{jt}, f_{jt}\right) \\
q(\mu) &= \prod_{k}^{K}\prod_{d}^{D} \mathcal{N}\left(\mu_{kd}|p_{kd}, q_{kd}^{-1}\right) \\
q\left(\tau\right) &= \prod_{k}^{K}\prod_{d}^{D} \mathcal{G}\left(\tau_{kd}|m_{kd}, s_{kd}\right)
\end{aligned} \tag{91}$$

By applying the mean field theory for the proposed HPYPGGM, we expand the ELBO as follows:

$$\begin{aligned}
\mathcal{L} = {} & E_q\left[\log p\left(X|I, W, \mu, \tau, \lambda\right)\right]) + E_q\left[\log p\left(I|\Phi'\right)\right] + E_q\left[\log p\left(\Phi'|\iota_a, \iota_b\right)\right] \\
& + E_q\left[\log p\left(W|p'\right)\right] + E_q\left[\log p\left(p'|\mathfrak{B}_a, \mathfrak{B}_b\right)\right] \\
& + E_q[\log p(\mu|p, q^{-1})] + E_q\left[\log p\left(\tau|m, s\right)\right] \\
& - E_q\left[\log q\left(W, \Phi', I, p', \mu, \tau, \lambda\right)\right]
\end{aligned} \tag{92}$$

where, E represents the expectation with respect to the subscripted parameter.

We obtain the updated equations for the variational parameters by maximizing Eq. (92) with respect to Eq. (91) as follows:

$$\varphi_{jtk} = \frac{\hat{\varphi}_{jtk}}{\sum_k^K \hat{\varphi}_{jtk}}, \quad \varrho_{jit} = \frac{\hat{\varrho}_{jit}}{\sum_t^T \hat{\varrho}_{jit}}$$

$$\hat{\varphi}_{jtk} = \exp\left\{ E_q\left[\log \Phi'_k\right] + \sum_{l=1}^{k-1} E_q\left[\log\left(1-\Phi'_l\right)\right] - \sum_i^N E_q\left[W_{jit}\right]\widetilde{R}\right\}$$

$$\hat{\varrho}_{jit} = \exp\left\{ E_q\left[\log p'_{jt}\right] + \sum_{s=1}^{t-1} E_q\left[\log\left(1-p'_{jt}\right)\right] - \sum_k^K E_q\left[I_{jtk}\right]\widetilde{R}\right\}\hat{\varphi}_{jtk}$$

$$\widetilde{R} = \sum_d^D E_q[\frac{1}{\lambda_{kd}}\log\tau_{kd} - \tau_{kd}|X_{jid}-\mu_{kd}|^{\lambda_{kd}}]$$

$$c_k = 1 - \gamma_{ak} + \sum_j^M \sum_t^T E_q\left[I_{jtk}\right]$$

$$d_k = \gamma_{bk} + k\gamma_{ak} + \sum_j^M \sum_t^T \sum_{l=k+1}^K E_q\left[I_{jtl}\right]$$

$$e_{jt} = 1 - \mathfrak{B}_{ajt} + \sum_i^N E_q\left[W_{jit}\right]$$

$$f_{jt} = \mathfrak{B}_{bjt} + t\mathfrak{B}_{ajt} + \sum_i^N \sum_{s=t+1}^T E_q\left[W_{jis}\right]$$

$$m_{kd} = \sum_j^M \sum_t^T \sum_i^N E_q[I_{jtk}]E_q[W_{jit}] + m_0 - 1$$

$$s_{kd} = \sum_j^M \sum_t^T \sum_i^N E_q[I_{jtk}]E_q[W_{jit}] + E_q[|X_{jid}-\mu_{kd}|^{\lambda_{kd}}] + s_0$$

$$p_{kd} = \frac{\frac{p_0 q_0}{2} + t_1}{q_{kd}}$$

$$q_{kd} = \frac{q_0}{2} + t_2$$

where m_0, s_0, p_0 and q_0 are the hyperparameters of m_{kd}, s_{kd}, p_{kd} and q_{kd}. t_1 and t_2 are defined as:

$$t_1 = \begin{cases} \sum_j^M \sum_t^T \sum_i^N \big((E_q[I_{jtk}]E_q[W_{jit}]E_q[\tau_{kd}]\frac{\lambda_{kd}}{4}(\lambda_{kd}-1)\mu_{kd}^{\lambda_{kd}-3}X_{jid}^2 + \\ (E_q[I_{jtk}]E_q[W_{jit}]E_q[\tau_{kd}]\frac{\lambda_{kd}}{2}\mu_{kd}^{\lambda_{kd}-2}X_{jid})\big), \text{if } X_{jid} < p_{kd} \\ \\ \sum_j^M \sum_t^T \sum_i^N E_q[I_{jtk}]E_q[W_{jit}]E_q[\tau_{kd}]\lambda_{kd}\frac{|X_{jid}|^{\lambda_{kd}}}{X_{jid}}, \text{otherwise} \end{cases} \tag{93}$$

$$t_2 = \begin{cases} \sum_j^M \sum_t^T \sum_i^N (E_q[I_{jtk}]E_q[W_{jit}]E_q[\tau_{kd}]p_{kd}^{\lambda_{kd}-2}), \text{if } X_{jid} < p_{kd} \\ \\ \sum_j^M \sum_t^T \sum_i^N (E_q[I_{jtk}]E_q[W_{jit}]E_q[\tau_{kd}]\frac{\lambda_{kd}}{2}(\lambda_{kd}-1)\frac{|X_{jid}^{\lambda_{kd}}|}{X_{jid}^2}), \text{otherwise} \end{cases}$$

The expected values for the equations in Eq. (93) are defined as follows:

$$\begin{aligned} & E_q\left[I_{jtk}\right] = \varphi_{jtk}, \quad E_q\left[W_{jit}\right] = \varrho_{jit} \\ & E_q\left[\log \Phi_k\right] = E_q\left[\log \Phi_k'\right] + \sum_{l=1}^{k-1} E_q\left[\log\left(1-\Phi_l'\right)\right] \\ & E_q\left[\log\left(\Phi_k'\right)\right] = \Psi\left(c_k\right) - \Psi\left(c_k + d_k\right) \\ & E_q\left[\log\left(1-\Phi_k'\right)\right] = \Psi\left(d_k\right) - \Psi\left(c_k + d_k\right) \\ & E_q\left[\log p_{jt}\right] = E_q\left[\log p_{jt}'\right] + \sum_{s=1}^{t-1} E_q\left[\log\left(1-p_{jt}'\right)\right] \\ & E_q\left[\log\left(p_{jt}'\right)\right] = \Psi\left(e_{jt}\right) - \Psi\left(e_{jt}+f_{jt}\right) \\ & E_q\left[\log\left(1-p_{jt}'\right)\right] = \Psi\left(f_{jt}\right) - \Psi\left(e_{jt}+f_{jt}\right) \end{aligned} \tag{94}$$

$$E_q[|X_{jid}-\mu_{kd}|^{\lambda_{kd}}] = \begin{cases} |X_{jid}|^{\lambda_{kd}} - \lambda_{kd}\frac{|X_{jid}|^{\lambda_{kd}}}{X_{jid}}p_{kd} + \\ \frac{\lambda_{kd}(\lambda_{kd}-1)}{2}\frac{|X_{jid}|^{\lambda_{kd}}}{X_{jid}^2}(\frac{1}{q_{kd}} + p_{kd}^2), \\ \text{if } X_{jid} > p_{kd} \\ \\ \mathbb{E}[|\mu_{kd}|^{\lambda_{kd}} - \lambda_{kd}\mu_{kd}^{\lambda_{kd}-1}X_{jid} + \\ \frac{\lambda_{kd}}{2}(\lambda_{kd}-1)\mu_{kd}^{\lambda_{kd}-2}X_{jid}^2], \\ \text{otherwise} \end{cases}$$

Using confluent hypergeometric function [47], the expected value of $|\mu_{kd}|^{\lambda_{kd}}$ can be

defined as :

$$E_q\left[|\mu_{kd}|^{\lambda_{kd}}\right] = \left(\frac{1}{\sqrt{q_{kd}}}\right)^{\lambda_{kd}} \cdot 2^{\lambda_{kd}/2} \frac{\Gamma\left(\frac{1+\lambda_{kd}}{2}\right)}{\sqrt{\pi}} \\ {}_1F_1\left(-\frac{\lambda_{kd}}{2}, \frac{1}{2}, -\frac{1}{2}\left(p_{kd}\right)^2 q_{kd}\right) \tag{95}$$

The shape parameter λ is given as follows [47]:

$$\lambda^{\star}_{kd} = \lambda_{kd} + \upsilon \Delta\lambda_{kd} \\ \text{where } \Delta\lambda_{kd} = -\frac{\mathcal{L}'_{kd}(q,\Theta)}{\mathcal{L}''_{kd}(q,\Theta)} \tag{96}$$

where υ is determined by the backtracking line search [54].

Algorithm 2 Hierarchical Pitman-Yor process of generalized Gaussian mixture model

1. **Initialization:** Set the truncation levels K and T.
2. Initialize the hyperparameters ι_a, ι_b, $\mathfrak{B}_a$, $\mathfrak{B}_b$, p_0, q_0, m_0 and s_0.
3. Initialize ϱ_{jit} using K-means
4. **loop**
   - i Estimate all the expected values in Eq. (94) and Eq. (95).
   - ii Update the parameters of the variational solution using the equations in Eq. (93).
   - iii Choose the step size υ by the backtracking line search and update λ_{kd} using Eq. (96).
   - iv The convergence criteria is reached when the difference between current and previous values of joint posteriors is less than $1e-9$.
5. **end**

4.3 Experimental results and discussion

4.3.1 Background subtraction

In this section, we employ the proposed HPYPGGM to address the problem of video background subtraction using a pixel-level evaluation approach [1]. This approach classifies whether the pixel belongs to the foreground or the background. Let us consider a frame $\mathcal{X}$ containing U pixels such that $\mathcal{X} = (\vec{X}_1, ..., \vec{X}_U)$. In the proposed algorithm, each pixel $\vec{X}_i$ represents red, green and blue (RGB) colors (3-dimensional) of the pixel which is modeled as a mixture of infinite GGD and the mixture components are shared between the groups (i.e., frames). The HPY process mixture

satisfies the above setting. We preprocess the frames by normalizing all the pixel values in an observed frame to unit sum. The preprocessed data is then used for learning the proposed HPYPGGM. In our mixture model, we can observe that some of the mixture components are used to model background pixels and the other models the foreground pixels. The final step in our framework is to determine if $\vec{X}_i$ is a foreground or a background pixel. In the proposed model, we assume a mixture component is classified as background if it occurs frequently, indicating high Φ and high precision τ [1]. We order the estimated components according to the product of $\Phi_k\tau_k$ and the resulting first B components are classified as background components, with B given by:

$$B = \arg\min_{b} \sum_{k=1}^{b} \Phi_k > \Upsilon \tag{97}$$

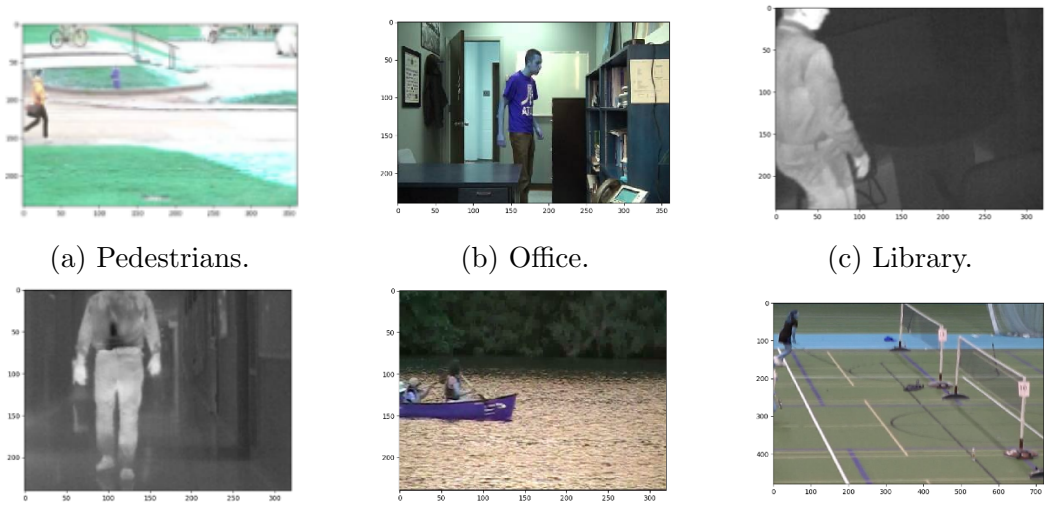
where Υ represent the minimum threshold of the data that should be accounted for the background in the frame, and the other components are classified as foreground components.

4.3.2 Results and discussion

In this section, we implement the proposed HPYPGGM algorithm on the challenging Change Detection dataset [70] which consists of 31 videos categorized into 6 different categories (baseline, shadows, dynamic background, intermittent object motion, camera jitter, and thermal). To evaluate the efficiency of the proposed model, we consider six videos of the Change Detection dataset which are described as follows:

- Pedestrians: This video sequence shows pedestrians walking in a park.
- Office: This video sequence shows a person walking around in an office.
- Library: This thermal video sequence shows a person walking in the library.
- Corridor: This thermal video sequence shows a person walking in the corridor.
- Canoe: This video sequence shows a moving canoe in a dynamic background.
- Badminton: This video sequence shows players playing badminton.

Sample images from the videos can be found in Fig. 11. In our experiments, we initialize threshold $\Upsilon = [0.55, 0.75]$ for different videos. Our results for each of the

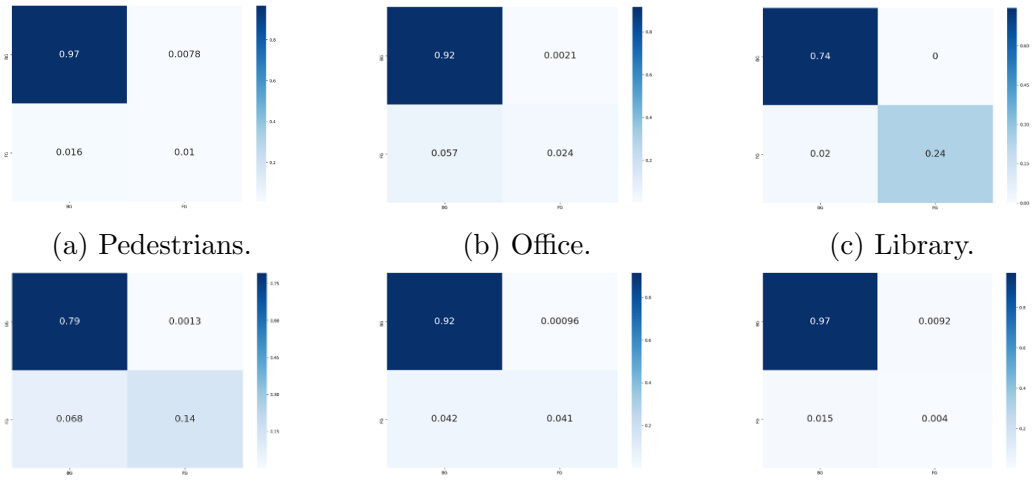


(a) Pedestrians. (b) Office. (c) Library.

(d) Corridor. (e) Canoe. (f) Badminton.

Figure 11: Sample frames of the video sequences from Change Detection dataset.

video sequences can be observed in confusion matrix form in Fig. 12. We evaluate the classification measure by accuracy, recall and precision which are defined as accuracy = (TP + TN)/Total no of observations, recall = TP/(TP + FN) and precision = TP/(TP + FP) where TP, TN, FP, and FN represent the total number of true positives, true negatives, false positives, and false negatives respectively. The reported results of precision and recall are based on the macro averages of the overall frames.



(a) Pedestrians. (b) Office. (c) Library.

(d) Corridor. (e) Caneo. (f) Badminton.

Figure 12: Confusion matrices of applying the proposed HPYPGGM model.

We compare our results with four other approaches from the literature; namely, K-means, GMM, variational GMM (VGMM) and Dirichlet process GMM (DPGMM). We set the number of components to 6 for K-means, GMM, VGMM and DPGMM. The threshold Υ is set to the same value as HPYPGGM for a fair comparison between the models. A visual comparison of the results from all the video sequences can be observed in Fig. 13 and Table 4 shows the comparison of the proposed HPYPGGM against K-means, GMM, VGMM and DPGMM. We can observe in the Pedestrians video sequence, that our model performed better in classifying between the background and foreground pixels while the others misclassified most of the background with foreground pixels. In the Office video sequence, K-means, GMM, VGMM and DPGMM were not able to precisely distinguish between the background and foreground pixels. This may be due to the close color intensity of the person's jeans with the color intensity of the box next to him. Nonetheless, HPYPGGM was able to segment a better foreground compared with the other models. All the models performed better in the Library video sequence where the background pixels are dark with perfect illumination, thereby resulting in a high accuracy with a supporting high precision and recall for all the models. In the Canoe video sequence all the models misclassified background water with foreground. However, HPYPGMM was able to give a better classification between the background and the foreground pixels. This can be observed clearly in Fig. 13. Similar results were obtained in the Badminton video sequence. Table 4 shows that the proposed HPYPGGM model outperformed K-means, GMM, VGMM and DPGMM in most cases in terms of precision, recall and accuracy.

Table 4: The macro average results of background subtraction with the Change Detection dataset.

Model	**Precision** (%)	**Recall** (%)	**Accuracy** (%)
Pedestrians			
K-means	67.30	73.11	96.48
GMM	60.12	73.42	94.02
VGMM	57.21	71.60	92.31
DPGMM	67.48	65.33	93.33
HPYPGGM	77.31	69.10	97.64
Office			
K-means	67.51	65.23	90.52
GMM	56.10	60.40	81.30
VGMM	58.92	65.11	81.51
DPGMM	89.48	87.12	87.21
HPYPGGM	93.31	65.20	94.05
Library			
K-means	98.97	96.02	98.04
GMM	98.56	96.01	98.06
VGMM	98.61	96.30	98.09
DPGMM	98.48	98.13	98.09
HPYPGGM	99.01	96.31	98.72
Corridor			
K-means	92.31	62.01	84.22
GMM	94.12	76.89	90.25
VGMM	95.23	78.12	83.50
DPGMM	94.48	93.21	93.01
HPYPGGM	96.01	83.51	93.43
Caneo			
K-means	80.10	78.26	93.22
GMM	76.82	77.12	92.59
VGMM	71.13	77.23	90.64
DPGMM	93.10	74.42	93.46
HPYPGGM	97.01	74.51	95.66
Badminton			
K-means	61.12	72.51	95.22
GMM	57.51	74.12	93.52
VGMM	57.12	74.71	93.50
DPGMM	66.48	65.47	94.73
HPYPGGM	66.01	66.52	97.40

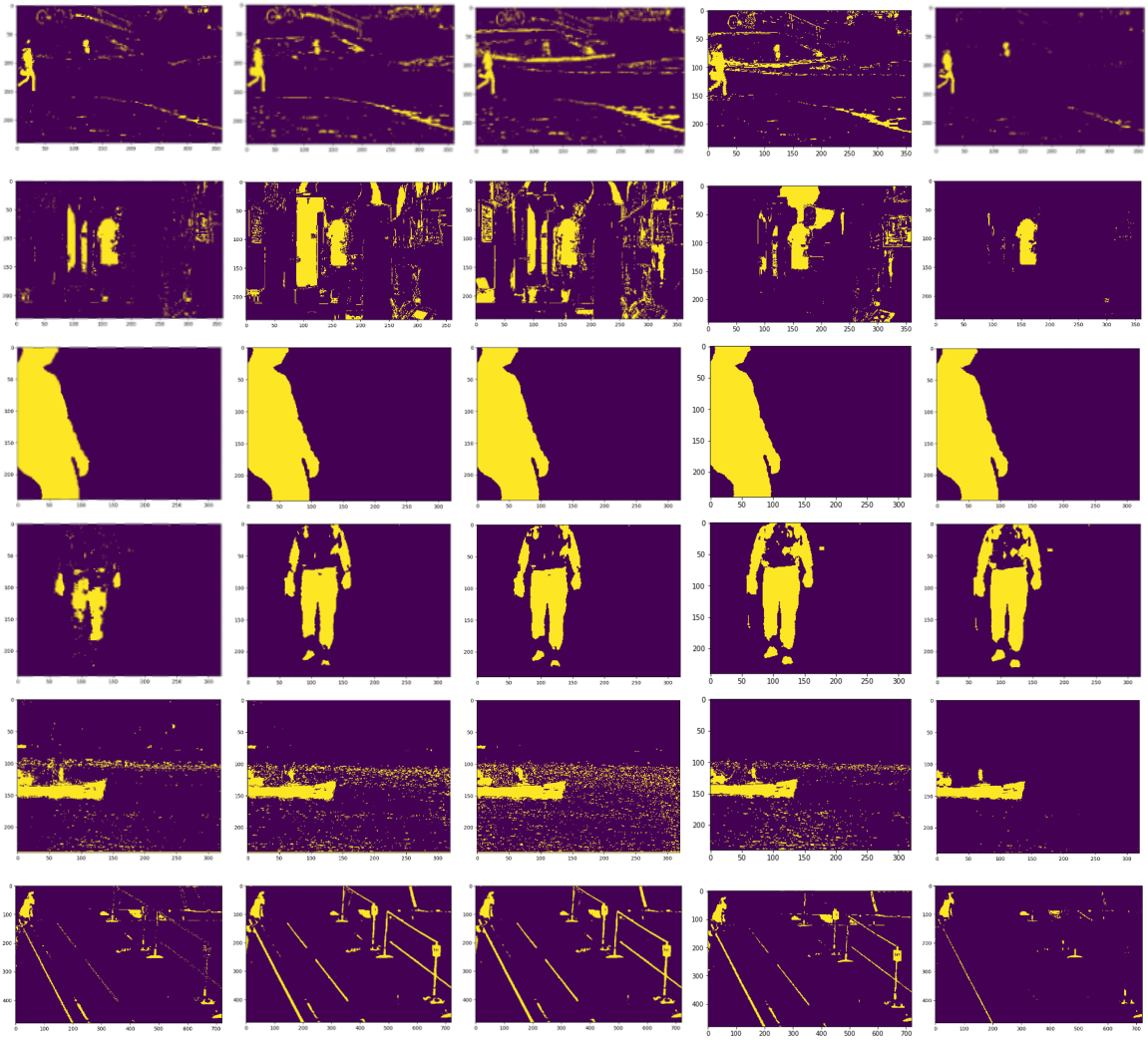


Figure 13: The foreground mask results for each of the original images (Pedestrians, Office, Library, Corridor, Caneo and Badminton from top to bottom respectively) obtained by K-means, GMM, VGMM, DPGMM and HPYPGMM algorithms are shown in columns 1 to 5 respectively.

Chapter 5

Conclusion

Clustering algorithms have been broadly applied in many research areas such as computer vision, signal processing, and pattern recognition. A mixture model, one of the most predominant statistical techniques, clusters data into a collection of homogeneous groups. Gaussian distribution has been widely used and studied with success for many applications involving computer vision, machine learning, image processing and statistical analysis. However, in many real applications, Gaussian fails to fit different shapes of data.

In this thesis, first we have presented a variational inference approach for generalized Gaussian Distribution. The algorithm was based on treating the shape parameter as a variable. Using the single-step update of Newton's method, the shape parameter is updated in the VM-step. Experimental results on medical, astrological, and image segmentation applications have shown the effectiveness of the algorithm when compared with the traditional models.

Second, we extended the variational inference approach to the infinite case using Dirichlet process and applied feature selection. Also, by extending the model to infinity with simultaneous feature selection, we were able to detect the number of mixture components and relevant features without the need to specify the number of mixture components a priori thereby resulting in an overall better accuracy. The variational learning approach aided in approximating the posteriors and experimental results have shown that the proposed variational IGGM with feature selection has favorable results compared to standard models.

Third, as an alternative to the proposed Dirichlet process prior, we considered

hierarchical Pitman-Yor process prior for our model. The proposed model tackles the estimation of parameters via variational learning. We inspected the benefits of our approach on video background subtraction using the challenging Change Detection dataset. The experimental results compared with the traditional models as well as the Dirichlet process of Gaussian distributions shows that our nonparametric Bayesian framework performed better over other models resulting in significant outcomes.

In conclusion, when compared with existing models and techniques which are mostly based on Gaussian assumption, our approach's can model non-Gaussian data with an efficient approximation of the model parameters resulting in better accuracy. Our models were also able to automatically determine the better number of mixture components resulting in overall better performance.

Future work could be dedicated to investigating online variational techniques for our proposed approaches and also extending the proposed models with component splitting. Another potential future work related to video background subtraction can be to modify the pixel based approach into segmenting each pixel of all the frames might lead to better over all performance in terms of classifying foreground and background pixels.

Bibliography

[1] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999.

[2] Chi Liu, Heng-Chao Li, Kun Fu, Fan Zhang, Mihai Datcu, and William J Emery. Bayesian estimation of generalized gamma mixture model based on variational em algorithm. *Pattern Recognition*, 87:269–284, 2019.

[3] Giancarlo Calvagno, Cristiano Ghirardi, Gian Antonio Mian, and Roberto Rinaldo. Modeling of subband image data for buffer control. *IEEE transactions on circuits and systems for video technology*, 7(2):402–408, 1997.

[4] Minh N Do and Martin Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE transactions on image processing*, 11(2):146–158, 2002.

[5] J-F Aujol, Gilles Aubert, and Laure Blanc-Féraud. Wavelet-based level set evolution for classification of textured images. *IEEE Transactions on Image Processing*, 12(12):1634–1641, 2003.

[6] Siu-Kai Choy and Chong-Sze Tong. Supervised texture classification using characteristic generalized gaussian density. *Journal of Mathematical Imaging and Vision*, 29(1):35–47, 2007.

[7] Mohand Said Allili, Nizar Bouguila, and Djemel Ziou. A robust video foreground segmentation by using generalized gaussian mixture modeling. In *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*, pages 503–509. IEEE, 2007.

[8] Mohand Saïd Allili, Nizar Bouguila, and Djemel Ziou. Finite general gaussian mixture modeling and application to image and video foreground segmentation. *Journal of Electronic Imaging*, 17(1):013005, 2008.

[9] Shu-Kai S Fan and Yen Lin. A fast estimation method for the generalized gaussian mixture distribution on complex images. *Computer Vision and Image Understanding*, 113(7):839–853, 2009.

[10] Thumpudi Naveen and John W Woods. Motion compensated multiresolution transmission of high definition video. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(1):29–41, 1994.

[11] Gabriele Moser, Josiane Zerubia, and Sebastiano B Serpico. Sar amplitude probability density function estimation based on a generalized gaussian model. *IEEE Transactions on Image Processing*, 15(6):1429–1442, 2006.

[12] Karnran Sharifi and Alberto Leon-Garcia. Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1):52–56, 1995.

[13] Pierre Moulin and Juan Liu. Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors. *IEEE Transactions on Information Theory*, 45(3):909–919, 1999.

[14] Thomas Fischer. A pyramid vector quantizer. *IEEE transactions on information theory*, 32(4):568–583, 1986.

[15] Keith A Birney and Thomas R Fischer. On the modeling of dct and subband image data for compression. *IEEE transactions on Image Processing*, 4(2):186–193, 1995.

[16] Charles Bouman and Ken Sauer. A generalized gaussian image model for edge-preserving map estimation. *ECE Technical Reports*, page 277, 1992.

[17] Yakoub Bazi, Lorenzo Bruzzone, and Farid Melgani. Image thresholding based on the em algorithm and the generalized gaussian distribution. *Pattern Recognition*, 40(2):619–634, 2007.

[18] Shu-Kai S Fan, Yen Lin, and Chia-Chan Wu. Image thresholding using a novel estimation method in generalized gaussian distribution mixture modeling. *Neurocomputing*, 72(1-3):500–512, 2008.

[19] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):674–693, 1989.

[20] Rajan L Joshi, Valerie J Crump, and Thomas R Fischer. Image subband coding using arithmetic coded trellis coded quantization. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6):515–523, 1995.

[21] Saeed Gazor and Wei Zhang. Speech probability distribution. *IEEE Signal Processing Letters*, 10(7):204–207, 2003.

[22] Kostas Kokkinakis and Asoke K Nandi. Exponent parameter estimation for generalized gaussian probability density functions with application to speech modeling. *Signal Processing*, 85(9):1852–1858, 2005.

[23] Demetrios Cantzos, Athanasios Mouchtaris, and Chris Kyriakakis. Multichannel audio resynthesis based on a generalized gaussian mixture model and cepstral smoothing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 215–218. IEEE, 2005.

[24] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[25] S Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE transactions on image processing*, 9(9):1532–1546, 2000.

[26] Pierre Moulin and Juan Liu. Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors. *IEEE transactions on Information Theory*, 45(3):909–919, 1999.

[27] Yakoub Bazi, Lorenzo Bruzzone, and Farid Melgani. Image thresholding based on the em algorithm and the generalized gaussian distribution. *Pattern Recognition*, 40(2):619–634, 2007.

[28] Minh N Do and Martin Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE transactions on image processing*, 11(2):146–158, 2002.

[29] Jacob Scharcanski. A wavelet-based approach for analyzing industrial stochastic textures with applications. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(1):10–22, 2006.

[30] Bruno Aiazzi, Luciano Alparone, and Stefano Baronti. Estimation based on entropy matching for generalized gaussian pdf modeling. *IEEE Signal Processing Letters*, 6(6):138–140, 1999.

[31] Mahesh K Varanasi and Behnaam Aazhang. Parametric generalized gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415, 1989.

[32] Minghong Pi. Improve maximum likelihood estimation for subband ggd parameters. *Pattern Recognition Letters*, 27(14):1710–1713, 2006.

[33] F Müller. Distribution shape of two-dimensional dct coefficients of natural images. *Electronics Letters*, 29(22):1935–1936, 1993.

[34] Sylvain Meignen and Hubert Meignen. On the modeling of small sample distributions with generalized gaussian density in a maximum likelihood framework. *IEEE Transactions on Image Processing*, 15(6):1647–1652, 2006.

[35] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[36] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[37] Ziyang Song, Samr Ali, and Nizar Bouguila. Bayesian learning of infinite asymmetric gaussian mixture models for background subtraction. In *International Conference on Image Analysis and Recognition*, pages 264–274. Springer, 2019.

[38] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[39] Eren Gultepe and Masoud Makrehchi. Improving clustering performance using independent component analysis and unsupervised feature learning. *Human-centric Computing and Information Sciences*, 2018.

[40] Nizar Bouguila and Djemel Ziou. A dirichlet process mixture of dirichlet distributions for classification and prediction. In *2008 IEEE workshop on machine learning for signal processing*, pages 297–302. IEEE, 2008.

[41] Michael Harrison and Arie Shirom. *Organizational diagnosis and assessment: Bridging theory and practice*. Sage Publications, 1998.

[42] Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166, 2004.

[43] Constantinos Constantinopoulos, Michalis K Titsias, and Aristidis Likas. Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):1013–1018, 2006.

[44] Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207, 2010.

[45] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[46] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[47] Srikanth Amudala, Samr Ali, Fatma Najar, and Nizar Bouguila. Variational inference of finite generalized gaussian mixture models. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2433–2439. IEEE, 2019.

[48] Srikanth Amudala, Samr Ali, and Nizar Bouguila. Variational inference of infinite generalized gaussian mixture models with feature selection. In *2020 IEEE International Conference on Systems, Man and Cybernetics (SMC) Organizing Committee*. IEEE, 2020.

[49] Srikanth Amudala, Samr Ali, and Nizar Bouguila. Background subtraction with a hierarchical pitman-yor process mixture model of generalized gaussian distribution. In *International Conference on Information Reuse and Integration*. IEEE, 2020.

[50] Tarek Elguebaly and Nizar Bouguila. Bayesian learning of finite generalized gaussian mixture models on images. *Signal Processing*, 91(4):801–820, 2011.

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

[52] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

[53] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[54] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[55] Linda Shapiro and C George. Stockman g: computer vision. In *Prentice Hall*. 2002.

[56] Lauren Barghout and Lawrence Lee. Perceptual information processing system, March 25 2004, US Patent App. 10/618,543.

[57] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[58] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[59] Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed variational dirichlet process mixture models. In *IJCAI*, volume 7, pages 2796–2801, 2007.

[60] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[61] Kamal Maanicshah, Nizar Bouguila, and Wentao Fan. Variational learning for finite generalized inverted dirichlet mixture models with a component splitting approach. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pages 1453–1458. IEEE, 2019.

[62] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[63] Teng Li, Tao Mei, In-So Kweon, and Xian-Sheng Hua. Contextual bag-of-words for visual categorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):381–392, 2010.

[64] 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.

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

[66] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[67] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[68] Deyang Wang, Weixin Xie, Jihong Pei, and Zongqing Lu. Moving area detection based on estimation of static background. *J Inform Comput Sci*, 2(1):129–134, 2005.

[69] Wentao Fan and Nizar Bouguila. Dynamic textures clustering using a hierarchical pitman-yor process mixture of dirichlet distributions. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 296–300. IEEE, 2015.

[70] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 387–394, 2014.