Variational Inference of Finite Generalized Gaussian Mixture Models

1st Srikanth Amudala
Concordia Institute for Information Systems Engineering
Concordia University
Montreal, Canada
srikanth.amudala@mail.concordia.ca

2nd Samr Ali
Department of Electrical and Computer Engineering
Concordia University
Montreal, Canada
al_samr@encs.concordia.ca

3rd Fatma Najar
Concordia Institute for Information Systems Engineering
Concordia University
Montreal, Canada
f_najar@encs.concordia.ca

4th Nizar Bouguila
Concordia Institute for Information Systems Engineering
Concordia University
Montreal, Canada
nizar.bouguila@concordia.ca

***Abstract*—This paper presents a variational inference method to analyze finite generalized Gaussian mixture models (GGMM) which incorporate several standard mixtures widely used in signal and image processing applications, such as Laplace and Gaussian. Our work is motivated by the fact that the generalized Gaussian distribution (GGD) can be applied on different types of data due to its shape flexibility. We present a method to evaluate the posterior distribution and Bayes estimators using variational expectation-maximization algorithm. The effective number of components of the GGMM is determined automatically. The experimental results demonstrate the effectiveness of the proposed algorithm by applying it to medical, astrological, and image segmentation applications; while comparing it to different other approaches.**

I. 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. However, in many real applications, Gaussian distribution fails to fit different shapes of data [1].

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 shape parameter. The GGD has three special cases with respect to the varying shape parameter namely the Laplacian, the Gaussian, and the asymptotically uniform distributions.

For instance, generalized Gaussian mixture model (GGMM) has been used in [2] for buffer control, in [3]–[5] for texture classification and retrieval, in [6]–[8] for video and image segmentation, in [9] for multiresolution transmission of high-definition video, in [10] for SAR images statistics modelling, in [11] for subband decomposition of video, in [12] for denoising applications, in [13], [14] for data and image compression, in [15] for edge modeling, in [16], [17] for image thresholding, in [18], [19] to fit subband histograms, in [20], [21] for speech modeling, and in [22] for multichannel audioresynthesis. Several methods have been proposed to estimate the parameters of GGMM such as entropy matching estimation [21], [23] and maximum likelihood estimation [3], [24]–[27] 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.

An alternative technique that has been gaining attention in the literature is the Bayesian method, for which a Markov chain Monte Carlo (MCMC) technique has been proposed in [28]. Although the MCMC algorithm yields better results for the inference of the GGMM, it is computationally intensive to evaluate the simulation-based estimator due to the Gibbs and Metropolis Hastings sampling.

In order to tackle problems related to both Bayesian and deterministic estimation, we propose in this paper 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. With the well defined prior distributions, the lower bound of the variational objective function is constructed. This facilitates the derivation of the closed-form updates in the variational expectation step (VE-step). Adopting the single-step update of Newton's method from [29], the closed-form updates for the power parameters are achieved in the variational maximization step (VM-step). By performing alternatively the VE-step and the VM-step, all the parameters of the GGMM

are updated. Experiments are performed based on medical, astrological, and image data sets to verify the effectiveness of the proposed method.

This paper is organized as follows. In Section 2, we present the variational inference of GGMM. In Section 3, we evaluate the performance of the proposed model on classification and segmentation applications. We conclude the paper in Section 4.

II. Variational Inference of the Generalized Gaussian Mixture Model

A. Generalized Gaussian Mixture Model

The one-dimensional GGMM for a variable $\mathcal{X} \epsilon R$ with parameters μ, τ, λ is defined as follows:

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

where $\tau = \left(\frac{1}{\sigma}\sqrt{\frac{\Gamma(\frac{3}{\lambda})}{\Gamma(\frac{1}{\lambda})}}\right)^{\lambda}$, $\Gamma(.)$ denotes the gamma function given by $\Gamma(z) = \int_0^{\infty} p^{z-1}e^{-p}dp$, where z and p are real variables. 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 larger the value, the flatter the probability density function. This means that λ determines the decay rate of the density function. Note that for the two special cases, when $\lambda = 2$ and $\lambda = 1$, the GGD is reduced to the Gaussian and the Laplacian distributions, respectively. If $\mathcal{X}$ follows a mixture of K GGDs, then

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

where $\pi_k (0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1)$ are the mixing weights and $p(\mathcal{X}|\mu_k,\tau_k,\lambda_k)$ is the GGMM likelihood of 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_i, let Z_i be K-dimensional vector known by the unobserved vector that assigns the appropriate mixture component X_i belongs to. Then, Z_{ik} is equal to 1 if X_i belongs to class k and 0, otherwise. Hence, the complete-data likelihood is given by:

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

The EM algorithm comprises of finding 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 k^{th} component of the mixture can be denoted as follows [30]:

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

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

1) Initialization of the mixture parameters.
2) E-step: Compute $\hat{Z}_{nk}^t$ (Eq. (6)) using the initialized parameters.
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 drawbacks, like convergence to local maxima due to its dependence on initialization. A detailed discussion of the disadvantages of the EM algorithm is in [31].

B. Variational Inference of the Generalized Gaussian Mixture Model

We propose a variational inference approach for the GGMM within the framework of Variational Expectation-Maximization (VEM) [32] [33] to achieve 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 and the approximate distribution q [33]. 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 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. With these priors, the posterior distributions for μ, τ, λ are defined as [30]:

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

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. 1:

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

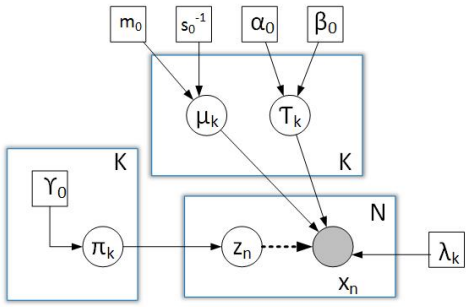


Fig. 1: Graphical model for the VGGM. The filled circle, unfilled circles and squares represent observations, random variables, and parameters, respectively. The dependency among the variables is represented by directional arrows.

Because of the nonlinearity of the shape parameter, the conjugate prior distribution can not be directly found. Therefore, we considered using the Taylor approximation to find 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 [1]. The conjugate exponential priors for μ and τ are Normal and Gamma distributions. Therefore, we specify all the priors according to:

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

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

We consider a variational distribution which factorizes into the latent variables and the parameters:

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

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

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

where $\mathbb{E}$ represents the expectation with respect to the subscripted parameter and $const$ denotes an additive constant. Substituting the two conditional distributions, and absorbing any terms that are independent of Z into the additive 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 optimal solution for the factor $q(Z)$ then follows the same functional form as the prior $p(Z|\pi)$. Note that because ρ_{nk} is given by the exponential of a real quantity, the quantities ρ_{nk} will be non-negative and will sum to one, as required. 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:

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

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 expression decomposes into a sum of terms with only π in addition to terms with μ and τ, which implies 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:

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

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

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

where γ has components γ_k that are given by:

$$\gamma_k = \gamma_0 + N_k \quad (26)$$

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

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

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

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

$$\begin{aligned} |\mu_k - X_n|^{\lambda_k} = &\mu_k^{\lambda_k} - \lambda_k \mu_k^{\lambda_k - 1} X_n + \\ &\frac{\lambda_k}{2}(\lambda_k - 1)\mu_k^{\lambda_k - 2} X_n^2 \end{aligned} \quad (29)$$

$if(X_n > \mu_k)$

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

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

$$m_k = \frac{\frac{s_0 m_0}{2} + p_1}{s_k} \quad (31)$$

$$s_k = \frac{s_0}{2} + p_2 \quad (32)$$

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:

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

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

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

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

Then using confluent hypergeometric function:

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

The following equation denotes the lower bound:

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

Given the posterior distributions from the VE-step, the VM-step updates the parameters by maximizing the approximate lower bound $\mathcal{L}$. To estimate the parameters of the GGMM (i.e. λ), the first-order derivative of the approximate lower bound is set to zero, leading to:

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

The second-order derivative is given by:

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

The shape parameter is now complete as:

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

where s is determined by the backtracking line search [34]. 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. (17)
6) Generate the responsibilities r_{nk} from Eq. (19)
7) Update $\alpha_k, \beta_k, \gamma_k \leftarrow$ from Eq. (34), Eq. (35) and Eq. (26)
8) Calculate m_k, s_k from Eq. (31), Eq. (32)
9) Choose the step size s by the backtracking line search
10) Update λ_k using Eq. (40)
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**

III. Experimental Results and Discussion

A. Implementation details

In the prior distributions for the shape parameter, 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 determine 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 accepted as the hyperparameter and parameter estimates in the Variational GGMM (VGGMM).

B. 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 with a positive heart disease. This database contains 76 attributes, but all distributed tests refer to employing a subset of 14. The objective field alludes to the presence of heart infection within the patient.

[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

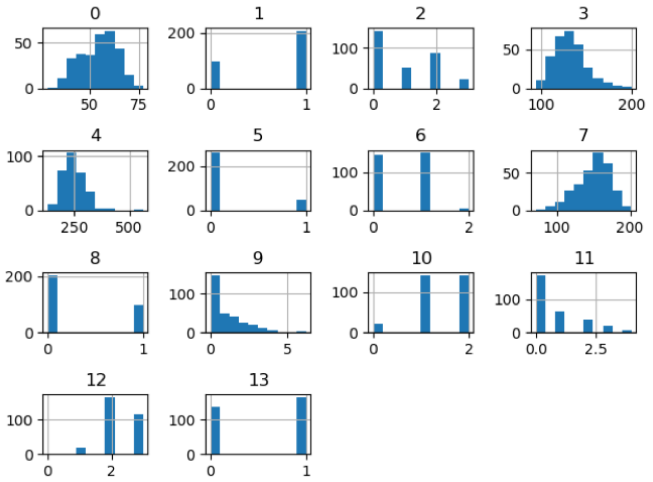


Fig. 2: Histograms of Heart Disease. Histogram-0 to Histogram-12 represent the features, Histogram-13 represents the target value. X-axis represents the value range and y-axis represents the frequency.

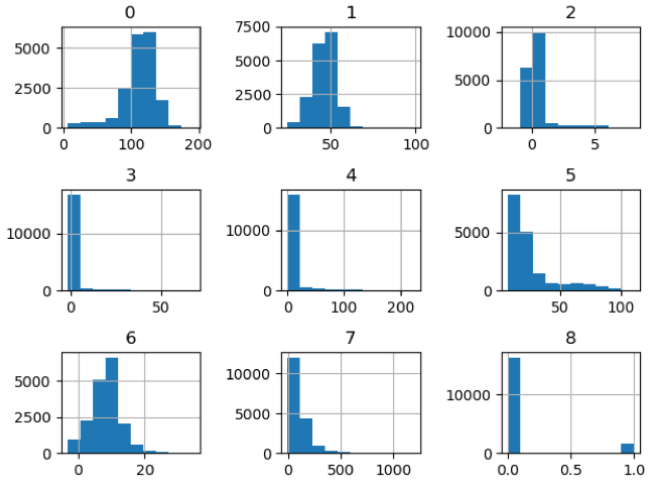
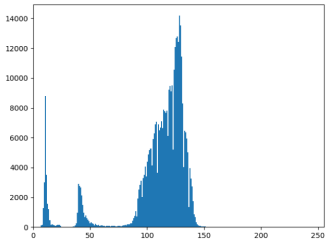


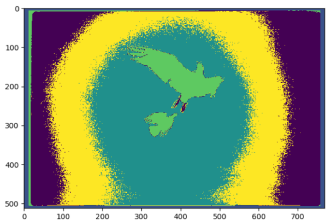
Fig. 3: Histograms of Pulsar Star. Histogram-0 to Histogram-7 represent the features, Histogram-8 represents the target value. X-axis represents the value range and Y-axis represents the frequency.

The second data set describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. Pulsars are a rare type of Neutron star that produce radio emission detectable here on earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. It has gained popularity over recent times to label the pulsar candidates to facilitate rapid analysis. Classification systems in particular are being widely adopted, which treat the candidate data sets as binary classification problems, which is a perfect fit for our comparison. The histograms of the input data sets have been presented in Fig. 2 and Fig. 3.

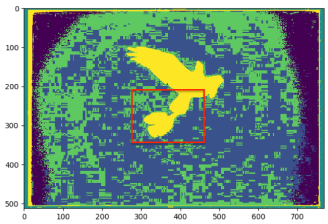
(a) Original Image



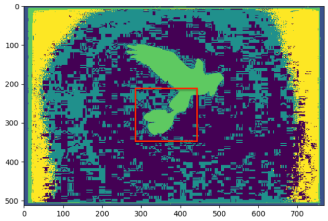
(b) histogram



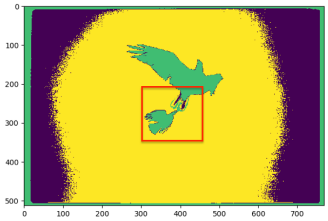
(c) K-means algorithm (K=5)



(d) GMM (K=5)



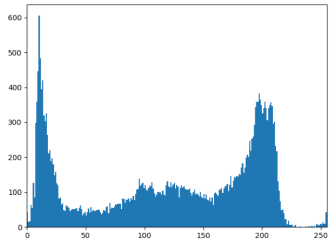
(e) VGMM (K=5)



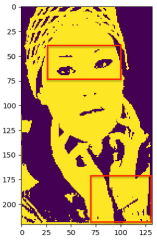
(f) VGGMM (K=5)

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

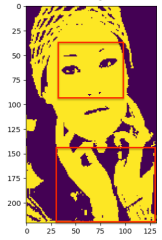
(a) Original Image



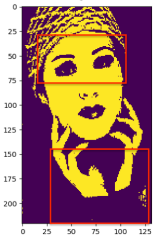
(b) histogram



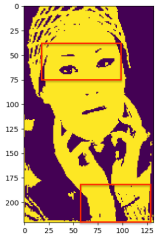
(c) K-means algorithm (K=2)



(d) GMM (K=2)



(e) VGMM (K=2)



(f) VGGMM (K=2)

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

We have implemented our VGGMM classifier using cross-validation with the split size of 4 for both the datasets. The label for each data point is determined with the largest component among the likelihood of the data point belonging to the classes. Table 1, presents the model accuracy in comparison with VGMM.

TABLE I: 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%

C. 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 [35] [36].

In the first experiment, we choose an image (768 × 512) with two birds in the sky to show the capability to segment small objects in a large background (Fig. 4a). The goal is to cluster the image into two classes: the objects with two birds and the sky. We set the number of components, $K = 5$. Comparing the results for K-means algorithm, GMM, and VGMM (Fig. 4c, Fig. 4d, Fig. 4e), there is a large misclassification of the sky and the space between the small object and the large object. Our method, VGGMM (Fig. 4f), is able to distinguish the two birds and to detect the components effectively. Compared 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 performed our evaluation on a human face image (132 × 221) as shown in Fig. 5a. The goal was 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), showed more details to offer more information for face recognition and image understanding.

IV. Conclusion

We have presented a variational inference approach for GGMM. The algorithm is based on treating the shape parameter as a variable. Subsequently, using Binomial Expansion with two cases, we estimate the expectation of the distributions. Hence, the posterior distributions of the inference can be updated by the corresponding hyperparameters. In the VM-step, the shape parameter is updated using the single-step update of the Newton's method.

Experimental results show that the VGGMM is an accurate model for medical, astrological, and image segmentation applications by effectively estimating the parameters. Moreover, in comparison with the VGMM results, the VGGMM has performed better for both classification and image segmentation.

References

[1] C. Liu, H.-C. Li, K. Fu, F. Zhang, M. Datcu, and W. J. Emery, "Bayesian estimation of generalized gamma mixture model based on variational em algorithm," *Pattern Recognition*, vol. 87, pp. 269–284, 2019.

[2] G. Calvagno, C. Ghirardi, G. A. Mian, and R. Rinaldo, "Modeling of subband image data for buffer control," *IEEE transactions on circuits and systems for video technology*, vol. 7, no. 2, pp. 402–408, 1997.

[3] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE transactions on image processing*, vol. 11, no. 2, pp. 146–158, 2002.

[4] J.-F. Aujol, G. Aubert, and L. Blanc-Féraud, "Wavelet-based level set evolution for classification of textured images," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1634–1641, 2003.

[5] S.-K. Choy and C.-S. Tong, "Supervised texture classification using characteristic generalized gaussian density," *Journal of Mathematical Imaging and Vision*, vol. 29, no. 1, pp. 35–47, 2007.

[6] M. S. Allili, N. Bouguila, and D. Ziou, "A robust video foreground segmentation by using generalized gaussian mixture modeling," in *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*. IEEE, 2007, pp. 503–509.

[7] ——, "Finite general gaussian mixture modeling and application to image and video foreground segmentation," *Journal of Electronic Imaging*, vol. 17, no. 1, p. 013005, 2008.

[8] S.-K. S. Fan and Y. Lin, "A fast estimation method for the generalized gaussian mixture distribution on complex images," *Computer Vision and Image Understanding*, vol. 113, no. 7, pp. 839–853, 2009.

[9] T. Naveen and J. W. Woods, "Motion compensated multiresolution transmission of high definition video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 1, pp. 29–41, 1994.

[10] G. Moser, J. Zerubia, and S. B. Serpico, "Sar amplitude probability density function estimation based on a generalized gaussian model," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1429–1442, 2006.

[11] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[12] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[13] T. Fischer, "A pyramid vector quantizer," *IEEE transactions on information theory*, vol. 32, no. 4, pp. 568–583, 1986.

[14] K. A. Birney and T. R. Fischer, "On the modeling of dct and subband image data for compression," *IEEE transactions on Image Processing*, vol. 4, no. 2, pp. 186–193, 1995.

[15] C. Bouman and K. Sauer, "A generalized gaussian image model for edge-preserving map estimation," *ECE Technical Reports*, p. 277, 1992.

[16] Y. Bazi, L. Bruzzone, and F. Melgani, "Image thresholding based on the em algorithm and the generalized gaussian distribution," *Pattern Recognition*, vol. 40, no. 2, pp. 619–634, 2007.

[17] S.-K. S. Fan, Y. Lin, and C.-C. Wu, "Image thresholding using a novel estimation method in generalized gaussian distribution mixture modeling," *Neurocomputing*, vol. 72, no. 1-3, pp. 500–512, 2008.

[18] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 674–693, 1989.

[19] R. L. Joshi, V. J. Crump, and T. R. Fischer, "Image subband coding using arithmetic coded trellis coded quantization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 515–523, 1995.

[20] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[21] K. Kokkinakis and A. K. Nandi, "Exponent parameter estimation for generalized gaussian probability density functions with application to speech modeling," *Signal Processing*, vol. 85, no. 9, pp. 1852–1858, 2005.

[22] D. Cantzos, A. Mouchtaris, and C. 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.* IEEE, 2005, pp. 215–218.

[23] B. Aiazzi, L. Alparone, and S. Baronti, "Estimation based on entropy matching for generalized gaussian pdf modeling," *IEEE Signal Processing Letters*, vol. 6, no. 6, pp. 138–140, 1999.

[24] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404–1415, 1989.

[25] M. Pi, "Improve maximum likelihood estimation for subband ggd parameters," *Pattern Recognition Letters*, vol. 27, no. 14, pp. 1710–1713, 2006.

[26] F. Müller, "Distribution shape of two-dimensional dct coefficients of natural images," *Electronics Letters*, vol. 29, no. 22, pp. 1935–1936, 1993.

[27] S. Meignen and H. Meignen, "On the modeling of small sample distributions with generalized gaussian density in a maximum likelihood framework," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1647–1652, 2006.

[28] T. Elguebaly and N. Bouguila, "Bayesian learning of finite generalized gaussian mixture models on images," *Signal Processing*, vol. 91, no. 4, pp. 801–820, 2011.

[29] C. Lemaréchal, "S. boyd, l. vandenberghe, convex optimization, cambridge university press, 2004 hardback, 65 uss, isbn 0 521 83378 7," 2006.

[30] T. Elguebaly and N. Bouguila, "Bayesian learning of finite generalized gaussian mixture models on images," *Signal Processing*, vol. 91, no. 4, pp. 801–820, 2011.

[31] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[32] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[33] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[34] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[35] L. Shapiro and C. George, "Stockman g: computer vision," in *Prentice Hall*, 2002.

[36] L. Barghout and L. Lee, "Perceptual information processing system," Mar. 25 2004, US Patent App. 10/618,543.