Entropy-based Variational Learning of Finite Inverted Beta-Liouville Mixture Model

Mohammad Sadegh Ahmadzadeh[1]
Narges Manouchehri[2],**Hafsa Ennajari**[2],**Nizar Bouguila**[2]
Manar Amayri[2,3],**Wentao Fan**[4]

[1]Department of Electrical Engineering, Concordia University, Canada
[2]Concordia Institute for Information Systems Engineering, Concordia University, Canada[2]
[3]Grenoble Institute of Technology, G-SCOP Lab, Grenoble, France
[4]Department of Computer Science and Technology, Huaqiao University, Xiamen, China
m_hmadza@encs.concordia.ca; narges.manouchehri@mail.concordia.ca
h_ennaja@encs.concordia.ca; nizar.bouguila@concordia.ca; manar.amayri@grenoble-inp.fr; fwt@hqu.edu.cn

Abstract

Mixture models are a common unsupervised learning technique that have been widely used to statistically approximate and analyse heterogenous data. In this paper, an effective mixture model-based approach for positive vectors clustering and modeling is proposed. Our mixture model is based on the inverted Beta-Liouville (IBL) distribution. To deploy the proposed model, we introduce an entropy-based variational inference algorithm. The performance of the proposed model is evaluated on two real-world applications, namely, human activity recognition and image categorization.

1 Introduction

Nowadays, we observe a rapid growth of complex data in all format due to the technological development. Thanks to the field of machine learning, we can automatically analyze and infer useful information from these large-scale data. In particular, data clustering is regarded as one of the most famous data analysis tools aiming at grouping data with similar patterns into the same cluster. Among existing clustering techniques, finite mixture models have shown great flexibility in data modeling and have been successfully used in various applications like pattern recognition and smart buildings (Nguyen et al. 2019), (Manouchehri et al. 2019a). Most existing related works assume that data samples are mainly drawn from a Gaussian distribution. However, this assumption has made the applicability of Gaussian mixture models very limited as this type of distribution is not suitable for all kinds of data (Elguebaly and Bouguila 2014). Therefore, researchers have proposed alternative non-Gaussian mixture models such as Dirichlet mixture models that are more recommended for proportional data modeling (Fan et al. 2017). In addition, inverted Dirichlet mixture models have shown great generalization capabilities due to their great modeling flexibility when dealing with asymmetric and symmetric semi-bounded data (Bdiri and Bouguila 2013; Mashrgy, Bdiri, and Bouguila 2014). However, a major drawback of the inverted Dirichlet mixture model is that it assumes that the feature vectors are positively correlated. Another

challenge that we face when working with mixture models is the parameters estimation that is usually not simple and analytically intractable. Multiple techniques have been proposed to overcome this problem, like maximum likelihood and expectation-maximization (EM) (Bishop 2006). Nonetheless, they are impractical as they tend to cause overfitting. On the other side Bayesian-based techniques such as Markov Chain Monte Carlo (MCMC) (Robert and Casella 2013) are more accurate but their convergence is not guaranteed. Another critical aspect to consider when designing a mixture model is the selection of the optimal number of components to represent the data. The minimum message length criterion (Bouguila and Ziou 2005) is an example of approaches that deals with this problem, to name a few.

In order to overcome the issues named above, we propose a novel finite variational inverted Beta-Liouville mixture model for data clustering. Our work is motivated by the success of the inverted Beta-Liouville (IBL) distribution (Maanicshah et al. 2020) (Fan and Bouguila 2016) in various applications. According to (Bouguila 2011) (Fan and Bouguila 2013) the Liouville family of distributions has revealed great modelling capabilities when dealing with data modeling. We also develop a variational inference algorithm that efficiently learns the latent parameters of our proposed model. The objective here is to find an approximate estimation of the true posterior that minimizes the Kullback-leibler divergence between the true posterior distribution and the approximated one. Furthermore, we propose an entropy-based variational learning algorithm to select the optimal number of mixture components. Initially, we start with one component, and continue incrementally to find the perfect number of components.

The remainder of this paper is organized as follows. In Section 2, we introduce the statistical background of the inverted Beta-Liouville mixture model. We describe the variational learning process of our model parameters in Section 3. Section 4 explains the entropy-based variational inference process algorithm for model selection. We provide the results of our experiments on two real-world datasets in Section 5. Finally, the paper is concluded in Section 6.

2 Model Specification

Let $\vec{X}_i = (X_{i1}, \ldots, X_{iD})$ be a D dimensional vector generated from a set of N independently identically distributed data samples $\mathcal{X} = (\vec{X}_1, \ldots, \vec{X}_N)$, drawn from an inverted Beta-Liouville distribution (Fan and Bouguila 2019):

$$p(\vec{X}_i \mid \alpha_{j1}, \ldots, \alpha_{jD}, \alpha_j, \beta_j) = \frac{\Gamma(\sum_{l=1}^D \alpha_{jl})\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}$$
$$\times \prod_{l=1}^D \frac{X_{il}^{\alpha_{jl}-1}}{\Gamma(\alpha_{jl})} \Big(\sum_{l=1}^D X_{il}\Big)^{\alpha_j - \sum_{l=1}^D \alpha_{jl}} \Big(1 + \sum_{l=1}^D X_{il}\Big)^{-(\alpha_j + \beta_j)} \tag{1}$$

The parameters of the probability density function for each component j are $\theta_j = (\alpha_{j1}, \ldots, \alpha_{jD}, \alpha_j, \beta_j)$. By assuming that each $\vec{X}_i$ is generated from a mixture of inverted Beta-Liouville distributions, we can define the mixture model as:

$$p(\mathcal{X} \mid \vec{\pi}, \Theta) = \prod_{i=1}^N \sum_{j=1}^M \pi_j p(\vec{X}_i \mid \theta_j) \tag{2}$$

where $p(\vec{X}_i \mid \theta_j)$ refers to the conditional probability of the data samples with respect to each component, $\Theta = (\theta_1, \ldots, \theta_M)$ and $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ is defined as the set of mixing coefficients with the constraints $\sum_{j=1}^M \pi_j = 1$ and $0 \le \pi_j \le 1$. Subsequently, we define an indicator matrix $\mathcal{Z} = (\vec{Z}_1, \ldots, \vec{Z}_N)$, where $\vec{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ is a binary latent vector associated with every data sample $\vec{X}_i$, with constraints $Z_{ij} \in \{0, 1\}$ and $\sum_{j=1}^M Z_{ij} = 1$. We assume that Z_{ij} is equal to 1 if $\vec{X}_i$ belongs to the component j, and zero otherwise. The conditional probability distribution of the indicator variable $\mathcal{Z}$ is given by:

$$p(\mathcal{Z} \mid \vec{\pi}) = \prod_{i=1}^N \prod_{j=1}^M \pi_j^{Z_{ij}} \tag{3}$$

From the equation above, we can define the following conditional distribution

$$p(\mathcal{X} \mid \mathcal{Z}, \Theta) = \prod_{i=1}^N \prod_{j=1}^M p(\vec{X}_i \mid \theta_j)^{Z_{ij}} \tag{4}$$

Since these parameters are positive, it would be convenient if we describe the priors with the Gamma distribution $\mathcal{G}(\cdot)$ as follows:

$$p(\alpha_{jl}) = \mathcal{G}(\alpha_{jl} \mid e_{jl}, f_{jl}) = \frac{f_{jl}^{e_{jl}}}{\Gamma(e_{jl})} \alpha_{jl}^{e_{jl}-1} e^{-f_{jl}\alpha_{jl}} \tag{5}$$

$$p(\beta_j) = \mathcal{G}(\beta_j \mid g_j, h_j) = \frac{h_j^{g_j}}{\Gamma(g_j)} \beta_j^{g_j-1} e^{-h_j\beta_j} \tag{6}$$

$$p(\alpha_j) = \mathcal{G}(\alpha_j \mid u_j, \nu_j) = \frac{\nu_j^{u_j}}{\Gamma(u_j)} \alpha_j^{u_j-1} e^{-\nu_j\alpha_j} \tag{7}$$

where all the hyperparameters are positive. At this point, we can represent the joint distribution for all the random variables as:

$$p(\mathcal{X}, \mathcal{Z}, \Theta \mid \vec{\pi}) = p(\mathcal{X} \mid \mathcal{Z}, \Theta)p(\mathcal{Z} \mid \vec{\pi})p(\vec{\alpha}_l)p(\vec{\beta})p(\vec{\alpha}) \tag{8}$$

$$p(\mathcal{X}, \mathcal{Z}, \Theta \mid \vec{\pi}) = \prod_{i=1}^N \prod_{j=1}^M \Big[\frac{\Gamma(\sum_{l=1}^D \alpha_{jl})\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}$$
$$\prod_{l=1}^D \frac{X_{il}^{\alpha_{jl}-1}}{\Gamma(\alpha_{jl})} \Big(\sum_{l=1}^D X_{il}\Big)^{\alpha_j - \sum_{l=1}^D \alpha_{jl}} \Big(1 + \sum_{l=1}^D X_{il}\Big)^{-(\alpha_j + \beta_j)} \Big]^{Z_{ij}}$$
$$\times \Big(\prod_{i=1}^N \prod_{j=1}^M \pi_j^{Z_{ij}}\Big) \prod_{j=1}^M \prod_{l=1}^D \Big[\frac{\nu_j^{u_j}}{\Gamma(u_j)} \alpha_j^{u_j-1} e^{-\nu_j\alpha_j}$$
$$\times \frac{h_j^{g_j}}{\Gamma(g_j)} \beta_j^{g_j-1} e^{-h_j\beta_j} \times \frac{f_{jl}^{e_{jl}}}{\Gamma(e_{jl})} \alpha_{jl}^{e_{jl}-1} e^{-f_{jl}\alpha_{jl}} \Big] \tag{9}$$

3 Model Learning with Entropy-Based Variational Inference

We define $Q(\Theta)$ as the approximation of the true posterior $p(\Theta \mid \mathcal{X}, \vec{\pi})$. The main goal of variational inference is to minimize the difference between the approximated distribution and the true posterior (Bishop 2006). The estimation of the true posterior distribution is accomplished with the Kullback-Leibler (KL) divergence between the two distributions. Therefore, the KL divergence between $p(\Theta \mid \mathcal{X}, \vec{\pi})$ and $Q(\Theta)$ is defined as follows:

$$KL(Q \| P) = \ln p(\mathcal{X} \mid \vec{\pi}) - \mathcal{L}(Q) \tag{10}$$

$$\mathcal{L}(Q) = \int Q(\Theta) \ln \Big(\frac{p(\mathcal{X}, \Theta \mid \vec{\pi})}{Q(\Theta)} \Big) d\Theta \tag{11}$$

According to the Jensen's inequality $\mathcal{L}(Q) \le \ln p(\mathcal{X} \mid \vec{\pi})$, $\mathcal{L}(Q)$ acts as the lower bound of the log likelihood. This means that we can minimize the KL divergence by maximizing the lower bound $\mathcal{L}(Q)$ (Bishop 2006). We adopt the mean field approximation approach in order to find the optimal parameters of the fully factorizable distribution Q, where $Q(\Theta) = Q(\mathcal{Z})Q(\vec{\alpha})Q(\vec{\beta})Q(\vec{\pi})Q(\vec{\alpha}_l)$. For a specific parameter $Q_s(\Theta_s)$, we can represent the optimal solution as:

$$Q_s(\Theta_s) = \frac{\exp\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \ne s}}{\int \exp\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \ne s} d\Theta} \tag{12}$$

where $\langle \cdot \rangle_{i \ne s}$ indicates the expectation with respect to all the parameters except Q_s. The variational approximations of our model are concluded as follows:

$$Q(\mathcal{Z}) = \prod_{i=1}^N \prod_{j=1}^M r_{ij}^{Z_{ij}} \,,\, r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^M \tilde{r}_{ij}} \tag{13}$$

$$Q(\vec{\alpha}) = \prod_{j=1}^M \mathcal{G}(\alpha_j \mid u_j^*, \nu_j^*) \,,\, Q(\vec{\beta}) = \prod_{j=1}^M \mathcal{G}(\beta_j \mid g_j^*, h_j^*) \tag{14}$$

$$Q(\vec{\alpha}_l) = \prod_{j=1}^M \prod_{l=1}^D \mathcal{G}(\alpha_{jl} \mid e_{jl}^*, f_{jl}^*) \tag{15}$$

$$\pi_j = \frac{1}{N} \sum_{i=1}^N r_{ij} \tag{16}$$

All needed quantities in above equations namely $\tilde{r}_{ij}$, (u_j^*, ν_j^*), (g_j^*, h_j^*), and (g_j^*, h_j^*) are developed in (Fan and Bouguila 2019).

The entropy-based variational Bayes for learning the finite inverted Beta-Liouville mixture model (EV-IBLMM) is developed in the same way as in (Fan et al. 2018) using differential entropy estimation and MeanNN entropy estimator.

Let $p(\vec{X}_i)$ be the probability density function of a random variable $\vec{X}_i = (\vec{X}_1, \ldots, \vec{X}_D)$ belonging to a set of N samples $\{\vec{X}_i, \ldots, \vec{X}_N\}$, $i = 1, \ldots, N$. The differential entropy of the continuous random variable $\vec{X}_i$ is defined by:

$$H\left(\vec{X}_i\right) = -\int p(\vec{X}_i) \log_2 p(\vec{X}_i) d\vec{X}_i \qquad (17)$$

The maximum differential entropy of the IBL is given by:

$$H_{IBL}\left[p\left(\vec{X}_i \mid \theta\right)\right] = \ln\left[\frac{\Gamma(\alpha)\Gamma(\beta)(\prod_{l=1}^{D}\Gamma(\alpha_l))}{\Gamma(\alpha+\beta)\Gamma(\sum_{l=1}^{D}\alpha_l)}\right]$$
$$+ (\alpha+\beta)(\psi(\beta) - \psi(\alpha+\beta)) + \sum_{l=1}^{D}\left[(1-\alpha_l)(\psi(\alpha_l)\right.$$
$$\left. - \psi\left(\sum_{l=1}^{D}\alpha_l\right)\right] + (D-\alpha)(\psi(\alpha) - \psi(\alpha+\beta)) \qquad (18)$$

Details about MeanNN entropy estimator can be found in (Faivishevsky and Goldberger 2009) and its application in our case is straightforward and similar to (Fan et al. 2018).

4 Experimental Results

In this section, we evaluate the performances of our proposed model EV-IBLMM based on real-world challenging datasets for human activity recognition and image categorization applications. We compare the results of our experiments with two other similar models, namely, Entropy-based Variational Dirichlet Mixture Model (EDMM) (Fan et al. 2017) and Entropy-based Multivariate Beta Mixture Model (EV-MBMM) (Manouchehri et al. 2019b).

Human Activity Recognition (HAR)

Human activity recognition in smart homes is a key factor to achieve home automation especially with the significant advancement in sensing technologies. It enables the smart applications to automatically react according to the human behaviour. However, automatically recognizing human activities like walking, sleeping and cooking is a challenging task, because human activities are complex by nature. In order to validate the performance of our model on the human activity recognition task, we used a dataset proposed in (van Kasteren, Englebienne, and Kröse 2011). This dataset was collected based on several types of wireless sensors including contact switches, pressure mats, and float sensors. It is to note that the recorded data are prone to noise because data might be lost if one of these sensors gets disconnected from the network. In this study, 20 sensors have been used, where each sensor represents a feature to our model. Since actions can overlap, the action that lasts longer

Table 1: Accuracy comparison of our EV-IBLMM approach and the baseline methods on the Human Activity Recognition dataset.

Method	Accuracy(%)
EV-IBLMM	**95.00**
EV-MBMM	93.30
EDMM	92.52

is maintained and kept. Thus, we consider a total of 6851 entries with 4 recorded activities. These activities include eating, sleeping, taking a shower and opening a door. The results of our proposed model and baselines are shown in Table 1. We can see that our model achieves the best accuracy performance among the other mixture models, which further demonstrates its efficiency for automatic human action recognition.

Image Categorization

Image categorization is considered as one of the important tasks of computer vision, and has witnessed much attention in the last decades. In this part of our experiments, we tested our proposed model on the image clustering task based on the Caltech101 image dataset (Fei-Fei, Fergus, and Perona 2004). This dataset contains images from 101 classes, with about 40-800 images in each category. For evaluation, we select a subset of 2033 data samples from 3 classes, namely, motorbikes, faces and airplanes. Some sample images from the three considered categories are illustrated in Figure 1. In
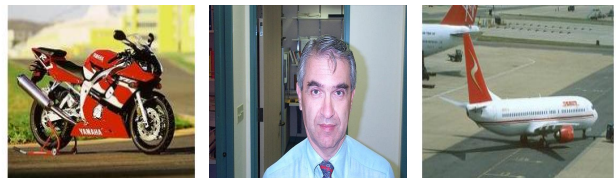

Figure 1: Sample images of each category from the considered subset of the Caltech101 dataset.

order to test our model on the Caltech101 dataset, we use SIFT (Lowe 2004) to extract the features of the designated images. This method has shown to be a good choice for this dataset. Then, we apply the K-means clustering algorithm on the results of the SIFT method, and use the output to create the Bag of Visual Words (BoVW) features. Table 2 illustrates the accuracy performance produced by each model. We observe that our EV-IBLMM model outperforms the EV-MBMM and EDMM models with a considerable margin of 1.8% and 3.2%, respectively, on the Caltech101 dataset. This highlights the effectiveness of our model in terms of model selection and data clustering.

5 Conclusion

In this paper, we proposed an unsupervised entropy-based variational method to learn the finite inverted Beta-Liouville mixture model. In order to select the optimal number of components, we used a novel entropy-based method for the

Table 2: Accuracy comparison of our EV-IBLMM approach and the baseline methods on the Caltech101 dataset.

Method	Accuracy(%)
EV-IBLMM	**90.20**
EV-MBMM	88.50
EDMM	87.10

splitting process. In this method, we conduct a comparison between the theoretical entropy and the entropy calculated with the MeanNN estimator. We proceed to split the component that has the highest weight difference into two new equal components, as we perceived that the mixture model is not giving a good description of that component. The experimental results demonstrate that our proposed model successfully outperforms two other competitive approaches on human activity recognition and image clustering tasks.

Acknowledgment

References

Bdiri, T., and Bouguila, N. 2013. Bayesian learning of inverted dirichlet mixtures for SVM kernels generation. *Neural Comput. Appl.* 23(5):1443–1458.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Bouguila, N., and Ziou, D. 2005. Mml-based approach for finite dirichlet mixture estimation and selection. In Perner, P., and Imiya, A., eds., *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, volume 3587 of *Lecture Notes in Computer Science*, 42–51. Springer.

Bouguila, N. 2011. Bayesian hybrid generative discriminative learning based on finite liouville mixture models. *Pattern Recognition* 44(6):1183–1200.

Elguebaly, T., and Bouguila, N. 2014. Background subtraction using finite mixtures of asymmetric gaussian distributions and shadow detection. *Mach. Vis. Appl.* 25(5):1145–1162.

Faivishevsky, L., and Goldberger, J. 2009. Ica based on a smooth estimation of the differential entropy. In *Advances in neural information processing systems*.

Fan, W., and Bouguila, N. 2013. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Rossi, F., ed., *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 1323–1329. IJCAI/AAAI.

Fan, W., and Bouguila, N. 2016. Model-based clustering based on variational learning of hierarchical infinite beta-liouville mixture models. *Neural Processing Letters* 44(2):431–449.

Fan, W., and Bouguila, N. 2019. Modeling and clustering positive vectors via nonparametric mixture models of liouville distributions. *IEEE Transactions on Neural Networks and Learning Systems*.

Fan, W.; Al-Osaimi, F. R.; Bouguila, N.; and Du, J. 2017. Proportional data modeling via entropy-based variational bayes learning of mixture models. *Applied Intelligence* 47(2):473–487.

Fan, W.; Bouguila, N.; Bourouis, S.; and Laalaoui, Y. 2018. Entropy-based variational bayes learning framework for data clustering. *IET Image Processing* 12(10):1762–1772.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. 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*, 178–178. IEEE.

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

Maanicshah, K.; Azam, M.; Nguyen, H.; Bouguila, N.; and Fan, W. 2020. Finite inverted beta-liouville mixture models with variational component splitting. In *Mixture Models and Applications*. Springer. 209–233.

Manouchehri, N.; Kalsi, J. S.; Amayri, M.; and Bouguila, N. 2019a. Finite two-dimensional beta mixture models: Model selection and applications. In *28th IEEE International Symposium on Industrial Electronics, ISIE 2019, Vancouver, BC, Canada, June 12-14, 2019*, 1407–1412. IEEE.

Manouchehri, N.; Rahmanpour, M.; Bouguila, N.; and Fan, W. 2019b. Learning of multivariate beta mixture models via entropy-based component splitting. In *IEEE Symposium Series on Computational Intelligence, SSCI 2019, Xiamen, China, December 6-9, 2019*, 2825–2832. IEEE.

Mashrgy, M. A.; Bdiri, T.; and Bouguila, N. 2014. Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. *Knowl. Based Syst.* 59:182–195.

Nguyen, H.; Rahmanpour, M.; Manouchehri, N.; Maanicshah, K.; Amayri, M.; and Bouguila, N. 2019. A statistical approach for unsupervised occupancy detection and estimation in smart buildings. In *2019 IEEE International Smart Cities Conference, ISC2 2019, Casablanca, Morocco, October 14-17, 2019*, 414–419. IEEE.

Robert, C., and Casella, G. 2013. *Monte Carlo statistical methods*. Springer Science & Business Media.

van Kasteren, T. L.; Englebienne, G.; and Kröse, B. J. 2011. Human activity recognition from wireless sensor network data: Benchmark and software. In *Activity recognition in pervasive intelligent environments*. Springer. 165–186.