Variational Learning for the Inverted Beta-Liouville Mixture Model and Its Application to Text Categorization

Yongfa Ling[1], Wenbo Guan[2], Qiang Ruan[3], Heping Song[4]*, Yuping Lai[5]

[1] School of Artificial Intelligence, Hezhou University, Hezhou (China)
[2] Institute of Information Engineering, CAS, Beijing (China)
[3] DigApis Information Security Technology Co., Ltd., Nantong, Jiangsu (China)
[4] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang (China)
[5] School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing (China)

Received 24 November 2021 | Accepted 22 June 2022 | Published 10 August 2022

Abstract

The finite invert Beta-Liouville mixture model (IBLMM) has recently gained some attention due to its positive data modeling capability. Under the conventional variational inference (VI) framework, the analytically tractable solution to the optimization of the variational posterior distribution cannot be obtained, since the variational object function involves evaluation of intractable moments. With the recently proposed extended variational inference (EVI) framework, a new function is proposed to replace the original variational object function in order to avoid intractable moment computation, so that the analytically tractable solution of the IBLMM can be derived in an effective way. The good performance of the proposed approach is demonstrated by experiments with both synthesized data and a real-world application namely text categorization.

Keywords

Bayesian Inference, Extended Variational Inference, Inverted Beta-Liouville Distribution, Mixture Model, Text Categorization.

DOI: 10.9781/ijimai.2022.08.006

I. Introduction

Positive data arise naturally in many real-world applications, such as object clustering [1], scene categorization [2], image segmentation [3], and object detection [4]. During the last decade, many non-Gaussian mixture models, e.g., the finite inverted Dirichlet mixture model (IDMM) [5], [6], the finite generalized inverted Dirichlet mixture model (GIDMM) [7], the finite generalized Gamma mixture model (GGaMM) [3] and the finite inverted Beta-Liouville mixture model (IBLMM) [8], were proposed to model and analyze positive data due to their powerful modeling capabilities. Among these mixture models, the IBLMM is one of the most popular approaches for modeling univariate and multivariate positive data. For example, the IBLMM is shown to be very flexible and powerful in analyzing and clustering text documents [8], therefore, modeling positive data with the IBLMM is well-motivated.

The major task in modeling the data with the finite mixture models is the learning of the model parameters, which refers to both estimating the model parameters and determining the number of components (i.e., the model complexity). A variety of approaches can be applied to address this problem, such as the expectation maximization (EM) algorithm [9], the Markov chain Monte Carlo (MCMC) [10], the expectation propagation (EP) [11] and the variational inference (VI) [12]. Among these approaches, the VI has been the most popular method. Much of its popularity is due to the fact that it may scale well to large applications. The main idea behind the VI is to find a approximate distribution for the intractable real posterior distribution by minimizing the Kullback-Leibler (KL) divergence of these two distributions. This is equivalent to maximizing the evidence lower bound (ELBO), which is also known as the variational objective function. Unfortunately, it is infeasible to obtain an analytical solution to the VI for many non-Gaussian mixtures, such as the IDMM, the GIDMM, the GGaMM and the IBLMM, since some computationally intractable moments exist in the ELBO. This problem can be solved by the recently proposed extended variational inference (EVI)[13]. The main idea behind the EVI framework is that the optimal solutions can be obtained by means of maximizing a lower bound of the ELBO. This bound can be obtained by introducing some tractable approximations to the original objective function.

Motivated by the powerful modeling capability of the IBLMM and the excellent performance achieved by the EVI framework, the EVI framework is applied to learn the IBLMM. The major contributions of this work can be summarized as follows. First, the analytical solution within the EVI framework for the IBLMM is derived. In this framework, the estimated values of all the involved parameters and the number of components can be simultaneously obtained. Second, the proposed approach is used in an important real-world application namely text categorization. Synthesized and real data evaluations demonstrate the good performance of the model trained by the proposed approach.

The reminder of this paper is organized as follows. In Section II, a brief review of the IBLMM is given. In Section III, the Bayesian learning algorithm with the EVI is derived. The experimental results on synthesized and real datasets are reported in Section IV. Finally, some conclusions are drawn in Section V.

* Corresponding author.

E-mail address: songhp@ujs.edu.cn (H. Song).

II. Preliminaries

A brief overview of the IBLMM is given first in this section. Then, a complete Bayesian framework for this model is presented.

A. Finite Inverted Beta-Liouville Mixture Model

If a D-dimensional random vector $\mathrm{x} = [x_1, \dots x_D]^T$ contains positive values, the underlying distribution of x can be modeled by the inverted Beta-Liouville (IBL) distribution. The probability density function (PDF) of the IBL distribution is given by [14]

$$p(\mathbf{x} \mid \boldsymbol{\alpha}, u, v) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \prod_{d=1}^{D} \frac{x_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times \left(\sum_{d=1}^{D} x_d\right)^{u - \sum_{d=1}^{D}\alpha_d} \left(1 + \sum_{d=1}^{D} x_d\right)^{-(u+v)} \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_D]^{\mathrm{T}}$, $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt$.

To model the multimodality of the observed data $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$, the mixture modeling technique [15] is used to construct the IBLMM with the PDF as follows

$$p(\mathbf{X} \mid \boldsymbol{\Lambda}, \mathbf{u}, \mathbf{v}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m \, p(\mathbf{x_n} \mid \boldsymbol{\alpha}_m, u_m, v_m) \tag{2}$$

where M is the number of components, $\boldsymbol{\pi} = [\pi_m, \dots, \pi_M]^{\mathrm{T}}$ is the mixing weights, $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M]$, $\boldsymbol{u} = [u_1, \dots, u_M]^{\mathrm{T}}$ and $\boldsymbol{v} = [v_1, \dots, v_M]^{\mathrm{T}}$ denote the parameter matrices.

B. Bayesian Framework for IBLMM

It is convenient to turn the mixture model in (2) into a latent variable model. For each vector $\boldsymbol{x}_n$, a latent vector variable $\boldsymbol{z}_n = [\mathrm{z}_{n1}, \dots, \mathrm{z}_{nM}]^{\mathrm{T}}$ is assigned, such that $\mathrm{z}_{nM} \in \{0,1\}$, $\sum_{m=1}^{M} \mathrm{z}_{nM} = 1$ and $\mathrm{z}_{nM} = 1$ if $\boldsymbol{x}_n$ is drawn from the mth component and 0 otherwise. Then, the latent variable model of IBLMM can be written as

$$p(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{m=1}^{M} \pi_m^{z_{nm}} \tag{3}$$

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Lambda}, \mathbf{u}, \mathbf{v}) = \prod_{n=1}^{N} \prod_{m=1}^{M} p\left(\mathbf{x}_n \mid \boldsymbol{\alpha}_m, u_m, v_m\right)^{z_{nm}} \tag{4}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_M]^{\mathrm{T}}$.

To formulate a full Bayesian mixture model, the conjugate priors on parameters $\boldsymbol{\Lambda}$, $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{\pi}$ have to be designated as follows:

$$p(\boldsymbol{\Lambda}) = \mathcal{G}(\boldsymbol{\Lambda} \mid \mathbf{g}, \mathbf{h}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{G}\left(\alpha_{md} \mid g_{md}, h_{md}\right) \tag{5}$$

$$p(\mathbf{u}) = \mathcal{G}(\mathbf{u} \mid \mathbf{s}, \mathbf{t}) = \prod_{m=1}^{M} \mathcal{G}\left(u_m \mid s_m, t_m\right) \tag{6}$$

$$p(\mathbf{v}) = \mathcal{G}(\mathbf{v} \mid \mathbf{p}, \mathbf{q}) = \prod_{m=1}^{M} \mathcal{G}\left(v_m \mid p_m, q_m\right) \tag{7}$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \mathbf{c}) = \frac{\Gamma(\sum_{m=1}^{M} c_m)}{\prod_{m=1}^{M} \Gamma\left(c_m\right)} \prod_{m=1}^{M} \pi_m^{c_m - 1} \tag{8}$$

where $\boldsymbol{g} = \{g_{md}\}$, $\boldsymbol{h} = \{h_{md}\}$, $\boldsymbol{s} = \{s_m\}$, $\boldsymbol{t} = \{t_m\}$, $\boldsymbol{p} = \{p_m\}$, $\boldsymbol{q} = \{q_m\}$, $\boldsymbol{c} = \{c_m\}$, $\mathcal{G}(\cdot)$ and $\mathrm{Dir}(\cdot)$ denote the Gamma distribution and the Dirichlet distribution, respectively.

Following the Bayes' theorem and combining (3), (4), (5), (6), (7) and (8), the joint distribution of the observation $\boldsymbol{X}$ and all the random variables $\boldsymbol{\Theta} = \{\boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\pi}\}$ is given by:

$$p(\mathbf{X}, \boldsymbol{\Theta}) = p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Lambda}, \mathbf{u}, \mathbf{v}) p(\mathbf{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\Theta}) p(\mathbf{u}) p(\mathbf{v}) \tag{9}$$

this equation has illustrated the relations of all the random variables entailed in the Baysian estimation of IBLMM. $\boldsymbol{Z}$ is the latent variable that indicates from which component the data is generated. π is the weight of each component. The other letters are the parameters of each IBLM.

III. Learning the Model

A. Extended Variational Inference

The VI framework [12] is commonly employed to estimate the parameters and determine the optimal number of components of the mixture models. The major goal is to find an approximate distribution $q(\boldsymbol{\Theta})$ for the true posterior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{X})$. The optimal $q(\boldsymbol{\Theta})$ can be obtained by maximizing the ELBO as follows:

$$\mathcal{L}(q) = \langle \ln p(\mathbf{X}, \boldsymbol{\Theta}) \rangle_q - \langle \ln q(\boldsymbol{\Theta}) \rangle_q \tag{10}$$

where $\langle \cdot \rangle_q$ denotes the expectation regarding the distribution q. Note that the $\mathcal{L}(q)$ is not analytically tractable for most of the non-Gaussian mixture models, such as the IDMM, the GIDMM, the GGaMM and the IBLMM, as (9) involves intractable moments. The recently proposed EVI framework [13] offers an effective way to proposed EVI framework [13] offers an effective way to framework is that if a "helping function" $\widetilde{p}\,(\mathbf{X\Theta})$, which satisfies the constraint $\mathrm{E}_q[\ln p(\mathbf{X}, \boldsymbol{\Theta})] \geq \mathrm{E}_q[\ln \widetilde{p}\,(\mathbf{X}, \boldsymbol{\Theta})]$, can satisfies the constraint $\mathrm{E}_q[\ln p(\mathbf{X}, \boldsymbol{\Theta})] \geq \mathrm{E}_q[\ln \widetilde{p}\,(\mathbf{X}, \boldsymbol{\Theta})]$, can be found, then the optimal solutions can be reached $\mathcal{L}(q)$. This bound is given by

$$\mathcal{L}(q) \geq \widetilde{\mathcal{L}}\,(q) = \mathrm{E}_q[\ln \widetilde{p}\,(\mathbf{X}, \boldsymbol{\Theta})] - \mathrm{E}_q[q(\boldsymbol{\Theta})] \tag{11}$$

To formulate a computationally tractable expression for the $\widetilde{\mathcal{L}}\,(q)$, the simplest approach called the mean-field approach is adopted which factorizes the $q(\boldsymbol{\Theta})$ as follows

$$q(\boldsymbol{\Theta}) = \prod_{n=1}^{N} \prod_{m=1}^{M} q\left(z_{nm}\right) \prod_{m=1}^{M} \prod_{d=1}^{D} q\left(\alpha_{md}\right) \times \prod_{m=1}^{M} [q(u_m) q(v_m) q(\pi_m)] \tag{12}$$

Then, the optimal form of $q(\Theta_k)$, denoted by $q^*(\Theta_k)$ in this case, is given by

$$\ln q_k^*(\Theta_k) = \langle \ln \widetilde{p}\,(\mathbf{X}, \Theta) \rangle_{s \neq k} + \mathrm{Cst} \tag{13}$$

where $\langle \cdot \rangle_{s \neq k}$ denotes the expectation regards all factors $q_s(\Theta_s)$ except for $s = k$ and "Cst" denotes a normalizing constant. In the EVI framework, all factors $q_s(\Theta_s)$ are need to be initiate first and then each factor is updated by updating the hyper-parameters.

B. Variational Distribution

This section details how (13) is applied to compute the variational factors. Note that the EVI is essentially iterative, since it represents a distribution factor applying knowledge about other factors. Following the principles of the EVI framework, the expectation of the joint distribution's logarithm is first calculated as

$$\begin{aligned}\langle \ln p(\mathbf{X},\mathbf{\Theta})\rangle &= \sum_{n=1}^{N}\sum_{m=1}^{M}\langle z_{nm}\rangle\{\langle \ln \pi_m\rangle + \mathcal{R}_m + \mathcal{F}_m \\ &+ \sum_{d=1}^{D}(\langle\alpha_{md}\rangle - 1)\ln x_{nd} + \ln\left(\sum_{d=1}^{D} x_{nd}\right)(\langle u_m\rangle \\ &- \sum_{d=1}^{D}\langle\alpha_{md}\rangle\Bigg) - (\langle u_m\rangle + \langle v_m\rangle)\ln\left(1+\sum_{d=1}^{D} x_{nd}\right)\Bigg\} \\ &+ \sum_{m=1}^{M}\sum_{d=1}^{D}[(g_{md}-1)\langle\ln\alpha_{md}\rangle - h_{md}\langle\alpha_{md}\rangle] \\ &+ \sum_{m=1}^{M}[(s_m-1)\langle\ln u_m\rangle - t_m\langle u_m\rangle] \\ &+ \sum_{m=1}^{M}[(p_m-1)\langle\ln v_m\rangle - q_m\langle v_m\rangle] \\ &+ \sum_{m=1}^{M}(c_m-1)\langle\ln\pi_m\rangle + \text{Cst}\end{aligned} \tag{14}$$

where $\mathcal{R}_m = \left\langle \ln\frac{\Gamma(\sum_{d=1}^{D}\alpha_{md})}{\prod_{d=1}^{D}\Gamma(\alpha_{md})}\right\rangle$, $\mathcal{F}_m = \left\langle \ln\frac{\Gamma(u_m+v_m)}{\Gamma(u_m)\Gamma(v_m)}\right\rangle$. It is noteworthy that (14) is not available in a closed form because it includes the intractable moments $\mathcal{R}_m, \mathcal{F}_m$. Following the principles of the aforementioned EVI framework, two "helping functions" $\widetilde{\mathcal{R}}_m, \widetilde{\mathcal{F}}_m$, satisfying $\mathcal{R}_m \geq \widetilde{\mathcal{R}}_m$, $\mathcal{F}_m \geq \widetilde{\mathcal{F}}_m$, respectively have to be found. According to [16], $\widetilde{\mathcal{R}}_m$ and $\widetilde{\mathcal{F}}_m$ are obtained as follows:

$$\begin{aligned}\widetilde{\mathcal{R}}_m &= \ln\frac{\Gamma(\sum_{d=1}^{D}\overline{\alpha}_{md})}{\prod_{d=1}^{D}\Gamma(\overline{\alpha}_{md})} + \sum_{d=1}^{D}\left[\Psi\left(\sum_{k=1}^{D}\overline{\alpha}_{mk}\right) - \Psi(\overline{\alpha}_{md})\right] \\ &\times [\langle\ln\alpha_{md}\rangle - \ln\overline{\alpha}_{md}]\overline{\alpha}_{md}\end{aligned} \tag{15}$$

$$\begin{aligned}\widetilde{\mathcal{F}}_m &= \ln\frac{\Gamma(\overline{u}_m+\overline{v}_m)}{\Gamma(\overline{u}_m)\Gamma(\overline{v}_m)} + [\Psi(\overline{u}_m+\overline{v}_m) - \Psi(\overline{u}_m)] \\ &\times (\langle\ln u_m\rangle - \ln\overline{u}_m)\overline{u}_m + [\Psi(\overline{u}_m+\overline{v}_m) - \Psi(\overline{v}_m)] \\ &\times (\langle\ln v_m\rangle - \ln\overline{v}_m)\overline{v}_m,\end{aligned} \tag{16}$$

where

$$\begin{aligned}\overline{\alpha}_{md} &= \langle\alpha_{md}\rangle, \overline{u}_m = \langle u_m\rangle \\ \overline{v}_m &= \langle v_m\rangle, \Psi(a) = \frac{\partial\ln\Gamma(a)}{\partial a}\end{aligned} \tag{17}$$

Insert (15) and (16) into (14) then a lower bound to $\langle \ln p(\boldsymbol{X},\boldsymbol{\Theta})\rangle$ is obtained as

$$\begin{aligned}\langle \ln\widetilde{p}(\mathbf{X},\mathbf{\Theta})\rangle &= \sum_{n=1}^{N}\sum_{m=1}^{M}\langle z_{nm}\rangle\{\langle \ln \pi_m\rangle + \widetilde{\mathcal{R}}_m + \widetilde{\mathcal{F}}_m \\ &+ \sum_{d=1}^{D}(\langle\alpha_{md}\rangle - 1)\ln x_{nd} + \ln\left(\sum_{d=1}^{D} x_{nd}\right) \\ &\times\left(\langle u_m\rangle - \sum_{d=1}^{D}\langle\alpha_{md}\rangle\right) - (\langle u_m\rangle + \langle v_m\rangle) \\ &\times\ln\left(1+\sum_{d=1}^{D} x_{nd}\right)\Bigg\} + \sum_{m=1}^{M}\sum_{d=1}^{D}[(g_{md}-1) \\ &\times\langle\ln\alpha_{md}\rangle - h_{md}\langle\alpha_{md}\rangle] \\ &+ \sum_{m=1}^{M}[(s_m-1)\langle\ln u_m\rangle - t_m\langle u_m\rangle] \\ &+ \sum_{m=1}^{M}[(p_m-1)\langle\ln v_m\rangle - q_m\langle v_m\rangle] \\ &+ \sum_{m=1}^{M}(c_m-1)\langle\ln\pi_m\rangle + \text{Cst.}\end{aligned} \tag{18}$$

Now, $\boldsymbol{\alpha}$, $\boldsymbol{u}$, and $\boldsymbol{v}$ are the i. i. d. variables. Details about solving the optimal variational factors using (13) is given as follows.

1. $q^*(\boldsymbol{Z})$: Including all terms that do not depend upon z_{nm} into a constant term, the equation (19) is obtained as follows

$$\ln q^*(z_{nm}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\ln\rho_{nm} + \text{Cst} \tag{19}$$

where

$$\begin{aligned}\ln\rho_{nm} &= \ln\pi_m + \widetilde{\mathcal{R}}_m + \widetilde{\mathcal{F}}_m + \sum_{d=1}^{D}(\overline{\alpha}_{md}-1)\ln x_{nd} \\ &+\left(\overline{u}_m - \sum_{d=1}^{D}\overline{\alpha}_{md}\right)\ln\left(\sum_{d=1}^{D} x_{nd}\right) \\ &-(\overline{u}_m + \overline{v}_m)\ln\left(1+\sum_{d=1}^{D} x_{nd}\right)\end{aligned} \tag{20}$$

Taking exponential of both sides of (19), $q^*(\boldsymbol{Z})$ is recognized to be a categorical density

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{m=1}^{M} r_{nm}^{z_{nm}} \tag{21}$$

where

$$r_{nm} = \frac{\rho_{nm}}{\sum_{m=1}^{M}\rho_{nm}} \tag{22}$$

where r_{nm} are nonnegative and have a unit sum.

2. $q^*(\boldsymbol{\Lambda})$: Absorbing any terms independent of $\boldsymbol{\alpha}_{md}$ into the additive constant results in

$$\ln q^*(\alpha_{md}) = (g^*_{md}-1)\ln\alpha_{md} - h^*_{md}\alpha_{md} + \text{Cst} \tag{23}$$

where g^*_{md} and h^*_{md} are defined by

$$g^*_{md} = g_{md} + \left[\Psi\left(\sum_{k=1}^{D}\overline{\alpha}_{md}\right) - \Psi(\overline{\alpha}_{md})\right]\overline{\alpha}_{md}\sum_{n=1}^{N}\langle z_{nm}\rangle \tag{24}$$

$$h^*_{md} = h_{md} - \sum_{n=1}^{N}\langle z_{nm}\rangle\left[\ln x_{nd} - \ln\left(\sum_{d=1}^{D} x_{nd}\right)\right] \tag{25}$$

Taking the exponential of both sides of (23), the equation (26) is obtained as follows

$$q^*(\mathbf{\Lambda}) = \prod_{m=1}^{M}\prod_{d=1}^{D}\mathcal{G}(\alpha_{md}\mid g^*_{md}, h^*_{md}) \tag{26}$$

3. $q^*(\boldsymbol{u})$: Any terms which are independent of u_m will be absorbed into the additive constant as

$$\ln q^*(u_m) = (s^*_m - 1)\ln u_m - t^*_{md}u_m + \text{Cst} \tag{27}$$

where s^*_m and t^*_m are given by

$$s^*_m = s_m + [\Psi(\overline{u}_m+\overline{v}_m) - \Psi(\overline{u}_m)]\overline{u}_m\sum_{n=1}^{N}\langle z_{nm}\rangle \tag{28}$$

$$t^*_m = t_m - \sum_{n=1}^{N}\langle z_{nm}\rangle\left[\ln\left(\sum_{d=1}^{D} x_{nd}\right) - \ln\left(1+\sum_{d=1}^{D} x_{nd}\right)\right] \tag{29}$$

Taking the exponential of both sides of (27), the equation (30) is obtained as follows

$$q(\mathbf{u}) = \prod_{m=1}^{M}\mathcal{G}(u_m\mid s^*_m, t^*_m) \tag{30}$$

4. $q^*(\boldsymbol{v})$: Considering the derivation of the update equation for the factor , the logarithm of the optimized factor is given by

$$\ln q^*(v_m) = (p_m^* - 1)\ln v_m - q_{md}^* v_m + \text{Cst} \tag{31}$$

where

$$p_m^* = p_m + [\Psi(\bar{u}_m + \bar{v}_m) - \Psi(\bar{v}_m)]\bar{v}_m \sum_{n=1}^{N} \langle z_{nm} \rangle \tag{32}$$

$$q_m^* = q_m + \sum_{n=1}^{N} \langle z_{nm} \rangle \ln\left(1 + \sum_{d=1}^{D} x_{nd}\right) \tag{33}$$

It is obvious that (31) has a similar form as to the logarithm of the Gamma prior density. Similarly, the equation (34) is obtained as follows

$$q^*(\mathbf{v}) = \prod_{m=1}^{M} \mathcal{G}(v_m \mid p_m^*, q_m^*) \tag{34}$$

5. $q^*(\boldsymbol{\pi})$: Keeping only terms that have a functional dependence on π_m, the equation (35) is obtained as follows

$$\ln q^*(\pi_m) = (c_m^* - 1)\ln \pi_m + \text{Cst} \tag{35}$$

where

$$c_m^* = \sum_{n=1}^{N} \langle z_{nm} \rangle + c_m \tag{36}$$

Taking the exponential of both sides of (35), the equation (37) is obtained as follows

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} \mid \mathbf{c}^*) = \frac{\Gamma(\sum_{m=1}^{M} c_m^*)}{\prod_{m=1}^{M} \Gamma(c_m^*)} \prod_{m=1}^{M} \pi_m^{c_m^* - 1} \tag{37}$$

All the expected values in the above equations are evaluated by

$$\bar{\alpha}_{md} = \frac{g_{md}^*}{h_{md}^*}, \langle \ln \alpha_{md} \rangle = \Psi(g_{md}^*) - \ln(h_{md}^*) \tag{38}$$

$$\bar{u}_m = \frac{s_m^*}{t_m^*}, \langle \ln u_m \rangle = \Psi(s_m^*) - \ln(t_m^*) \tag{39}$$

$$\bar{v}_m = \frac{p_m^*}{q_m^*}, \langle \ln v_m \rangle = \Psi(p_m^*) - \ln(q_m^*) \tag{40}$$

$$\langle z_{nm} \rangle = r_{nm}, \langle \pi_m \rangle = \frac{c_m^*}{\sum_{m=1}^{M} c_m^*}$$
$$\langle \ln \pi_m \rangle = \Psi(c_m^*) - \Psi\left(\sum_{m=1}^{M} c_m^*\right) \tag{41}$$

C. Full Variational Learning Algorithm

With the above obtained variational factors in hand, it is straightforward to evaluate the lower bound (11) for this model. In practice, it is useful to be able to monitor the bound during the re-estimation in order to test for convergence. The lower bound (11) is given by

$$\widetilde{\mathcal{L}}(q) = \langle \ln \widetilde{p}(\mathbf{X}, \boldsymbol{\Theta}) \rangle - \langle \ln q^*(\mathbf{Z}) \rangle - \langle \ln q^*(\boldsymbol{\Lambda}) \rangle$$
$$- \langle \ln q^*(\mathbf{u}) \rangle - \langle \ln q^*(\mathbf{v}) \rangle - \langle \ln q^*(\boldsymbol{\pi}) \rangle \tag{42}$$

where $\langle \ln \widetilde{p}(\mathbf{X}, \boldsymbol{\Theta}) \rangle$ is computed using (18). The other terms in the bound are easily evaluated to give the following results:

$$\langle \ln q^*(\mathbf{Z}) \rangle = \sum_{n=1}^{N} \sum_{m=1}^{M} r_{nm} \ln r_{nm} \tag{43}$$

$$\langle \ln q^*(\boldsymbol{\Lambda}) \rangle = \sum_{m=1}^{M} \sum_{d=1}^{D} [g_{md}^* \ln h_{md}^* - \ln \Gamma(g_{md}^*)$$
$$+ (g_{md}^* - 1)\langle \ln \alpha_{md} \rangle - h_{md}^* \langle \alpha_{md} \rangle] \tag{44}$$

$$\langle \ln q^*(\mathbf{u}) \rangle = \sum_{m=1}^{M} [s_m^* \ln t_m^* - \ln \Gamma(s_m^*)$$
$$+ (s_m^* - 1)\langle \ln u_m \rangle - t_m^* \langle u_m \rangle] \tag{45}$$

$$\langle \ln q^*(\mathbf{v}) \rangle = \sum_{m=1}^{M} [p_m^* \ln q_m^* - \ln \Gamma(p_m^*)$$
$$+ (p_m^* - 1)\langle \ln v_m \rangle - q_m^* \langle v_m \rangle] \tag{46}$$

$$\langle \ln q^*(\boldsymbol{\pi}) \rangle = \ln \frac{\Gamma(\sum_{m=1}^{M} c_m^*)}{\prod_{m=1}^{M} \Gamma(c_m^*)} + \sum_{m=1}^{M} (c_m^* - 1)\langle \ln \pi_m \rangle \tag{47}$$

The analytically tractable solution for Bayesian estimation of the IBLMM can be obtained in a similar way to the conventional EM algorithm. This inference algorithm is summarized in the Algorithm 1.

Algorithm 1. Algorithm for EVI-based Bayesian IBLMM

1. Set the initial values of M, g_{md}, h_{md}, s_m, t_m, p_m, q_m, c_m.
2. Initialize r_{nm} by K-Means algorithm
3. **repeat**
4. The variational E-step: Update $q^*(\boldsymbol{Z})$ according to (21).
5. The variational M-step: Update $q^*(\boldsymbol{\Lambda})$, $q^*(\boldsymbol{u})$, $q^*(\boldsymbol{v})$ and $q^*(\boldsymbol{\pi})$ according to (26), (30), (34), and (37), respectively.
6. **until** Stop criterion is reached.
7. Determine the best number of components M via annihilating the components with mixing weights $\pi_m \le 10^{-5}$.

IV. Experiments and Results

In this section, the proposed variational method refered to as EVI-IBLMM is validated through both synthesized datasets and real datasets. The goal of the synthesized dataset validation is to investigate the accuracy of the EVI-IBLMM algorithm in terms of parameter estimation and model selection. The goal of the real dataset validation is to compare the EVI-IBLMM to three other methods: the IDMM applying the EVI technique (EVI-IDMM) [6], the GIDMM applying the EVI technique (EVI-GIDMM) [13] and the GaMM applying the EVI technique (EVI-GaMM) [4]. To provide broad noninformative prior distributions, we set the hyperparameters of the prior distribution as $g_{md} = s_m = p_m = 1$, $h_{md} = t_m = q_m = 0.1$, $c_m = 0.001$, and initialize the number of components with large value (15 in this paper). The initial values of r_{nm} are obtained using the K-means algorithm. Note that this specific selection was based on our experiments and was found to be convenient and effective in our case. When the EVI-IBLMM algorithm stops, the posterior means are taken as the parameter estimates in the IBLMM.

A. Synthesized Data Validation

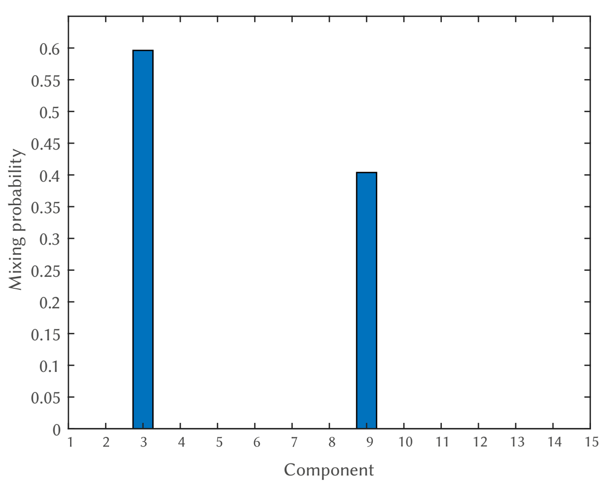
The performance of the proposed EVI-IBLMM in terms of estimation and determination through quantitative analysis on four 2-D synthesized datasets is first evaluated, which are generated from four known IBLMMs with different parameters. It is worth noting that the selection of $D = 2$ is purely for ease of representation. Table I shows the actual parameters for the four IBLMMs. The initial number of components for each dataset are set to double amounts of the actual number of components with equal mixture weights. The average estimated parameters of the four generated datasets over 20 runs of

TABLE I. True Values of the Parameters in the IBLMM Applied to Generate the Four Synthesized Datasets

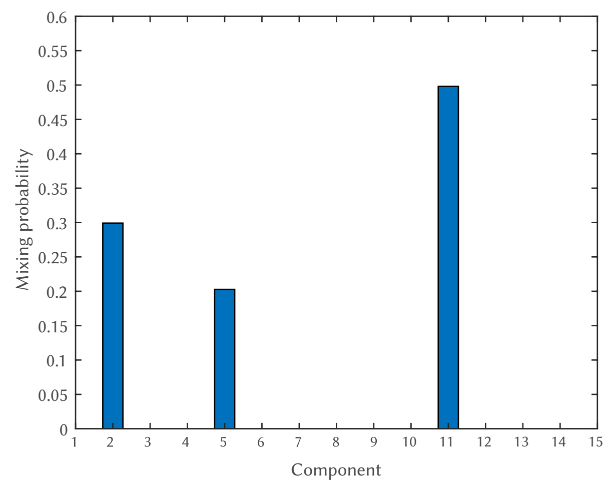
Dataset	m	α_{m1}	α_{m2}	u_m	v_m	π_m
A	1	12.00	24.00	8.50	12.50	0.400
	2	21.00	15.00	18.00	5.00	0.600
B	1	12.00	24.00	8.50	12.50	0.200
	2	21.00	15.00	18.00	5.00	0.300
	3	18.50	8.00	4.00	16.50	0.500
C	1	12.00	21.00	8.50	12.50	0.100
	2	21.00	35.00	18.00	5.00	0.200
	3	32.00	28.00	4.00	16.50	0.300
	4	2.00	18.00	24.00	8.00	0.400
D	1	21.00	6.00	18.00	24.00	0.100
	2	2.00	28.00	8.00	15.00	0.200
	3	18.00	68.00	24.00	16.00	0.250
	4	76.00	8.00	4.00	18.00	0.300
	5	2.00	4.00	4.00	12.00	0.150

TABLE II. The Mean of the Estimated Parameters for the Synthesized Datasets Over 20 Runs of the EVI-IBLMM Algorithm

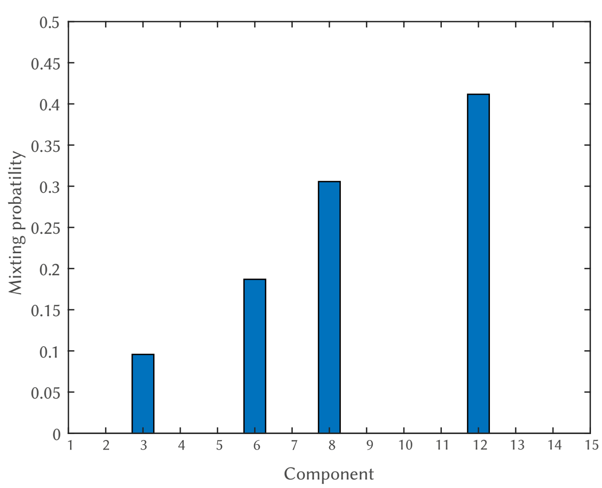
Dataset	N_m	m	$\hat{\alpha}_{m1}$	$\hat{\alpha}_{m2}$	$\hat{u}_m$	$\hat{v}_m$	$\hat{\pi}_m$
A	200	1	11.99	23.95	8.56	12.51	0.400
	300	2	21.27	15.20	18.10	5.00	0.600
B	120	1	11.31	22.59	8.50	12.54	0.200
	180	2	20.81	14.93	18.50	5.13	0.300
	300	3	18.30	8.01	4.18	17.09	0.500
C	80	1	12.46	21.64	9.20	14.12	0.098
	160	2	19.84	33.52	18.30	5.08	0.202
	240	3	30.68	26.81	4.07	16.76	0.300
	320	4	2.00	18.12	24.32	8.21	0.400
D	100	1	22.26	6.42	17.70	23.46	0.103
	200	2	1.98	27.09	7.80	15.03	0.201
	250	3	16.61	64.69	23.79	15.82	0.253
	300	4	73.02	7.48	4.04	18.10	0.302
	150	5	2.32	4.14	3.98	12.11	0.141



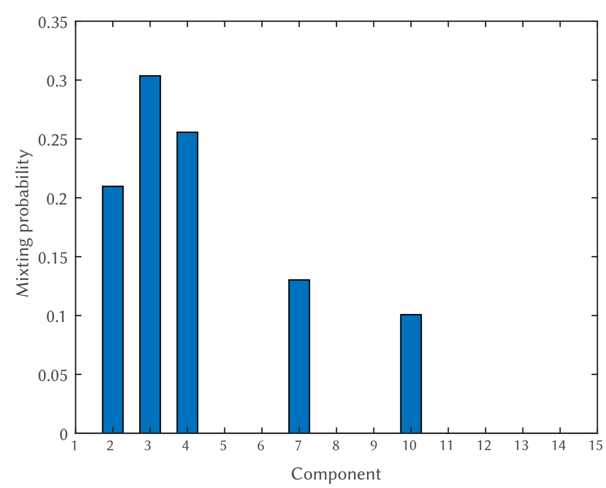
(a) Dataset A



(b) Dataset B

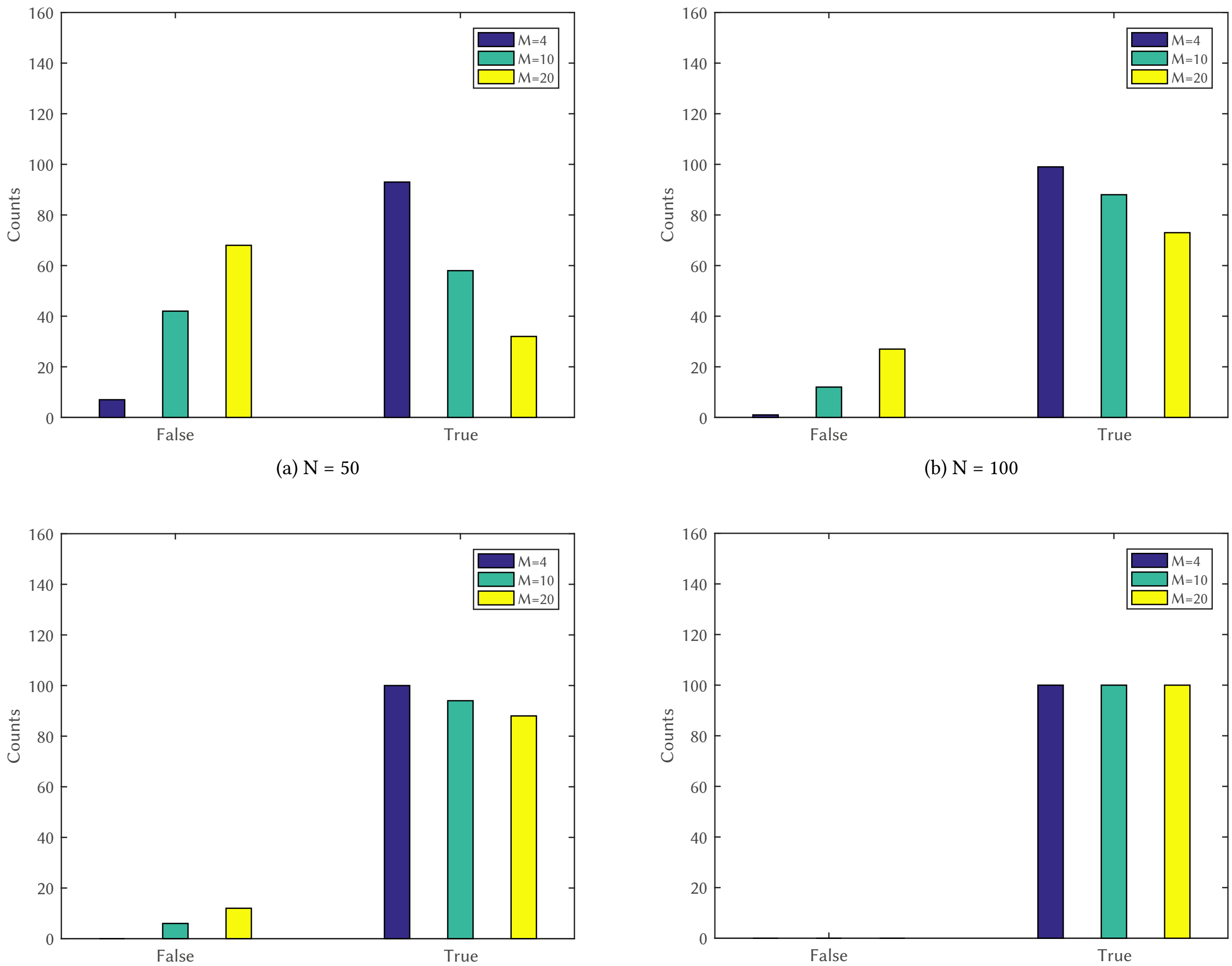


(c) Dataset C



(d) Dataset D

Fig. 1. Estimated mixing probabilities of components for the synthesized datasets. (a) Dataset A. (b) Dataset B. (c) Dataset C. (d) Dataset D.



(a) N = 50

(b) N = 100

(c) N = 200

(d) N = 500

Fig. 2. The counts of the estimated number of components over 100 runs of simulations based on dataset A. M denotes the initial number of components and N denotes the sample size.

simulations are reported in Table II. According to these results, the proposed EVI-IBLMM algorithm is capable of accurately estimating both the parameters and the mixing weights of the IBLMM. Next, the model selection capability of the EVI-IBLMM algorithm is investigated. When the initial number of components is larger than the true one, the EVI-IBLMM algorithm is capable of forcing some of the mixing weights to approach zero. These components make little contribution to the model, thus they can be eliminated. The EVI-IBLMM algorithm is initiated with a mixture of many components (15 in this paper) and equal mixture weights. Fig. 1 shows the estimated mixture weights of each component for the different generated datasets after convergence. According to these results, it can be clearly observed that the EVI-IBLMM algorithm is able to effectively determine the model complexity. Then, the effect of initial number of components upon the resulting model complexity is investigated. Based on dataset A, Fig. 2 shows the effect of initial number of components on the resulting model complexity over 100 runs of simulations. In Fig. 2, “True” donates that the model has correctly converged to the initial number of components and “False” means that the model does not have the same componets number with the initial ones after training. According to the results shown in this picture, the EVI-IBLMM algorithm is capable of identifying the accurate number of components regardless of whether the sample size is small or large. Moreover, as the sample size gets larger, the effect of the initial number of components gets more insignificant. Finally, the convergence of the EVI-IBLMM algorithm is investigated. Fig. 3 shows the value of the variational objective function in each iteration. According to this figure, it is clear that the variational objective function is always increasing during iterations, thus the convergence is demonstrated.

B. Text Categorization

Text categorization refers to the task of automatically assigning unlabeled text documents into predefined categories. During the past few decades, this task has attracted considerable attention from researchers due to many reasons, such as the hug amount of digital documents that are easily available and the increasing demand to organize, store, and retrieve these documents accurately and efficiently. Efficient text categorization are beneficial for many applications, such as document processing and visualization [17], digital information search [18], and information retrieval [19]. This problem is challenging and different statistical methods were proposed and applied in the past. Although different, most of the proposed techniques addressed this problem as following: First, a set of labeled text documents which belong to a certain number of classes are given to train the model. In our experiment, the data that has the same label is used to train one IBLMM, after training, the number of IBLMM is equal to the number of categories; Second, a new unobserved text is assigned to the category with the highest similarity regarding its content by the model.

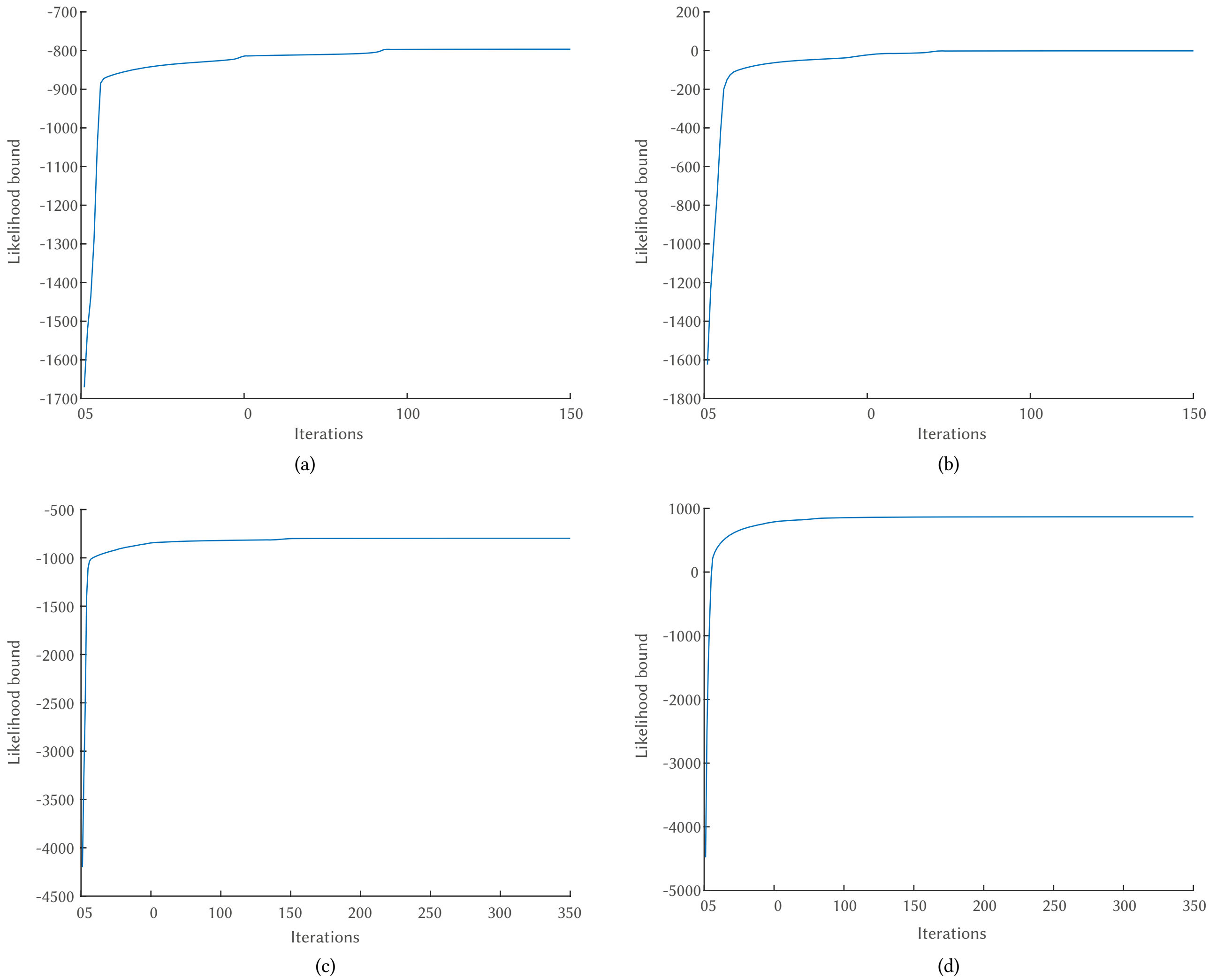


Fig. 3. Convergence of the proposed EVI-IBLMM algorithm for the different synthesized datasets. (a) Dataset A. (b) Dataset B. (c) Dataset C. (d) Dataset D.

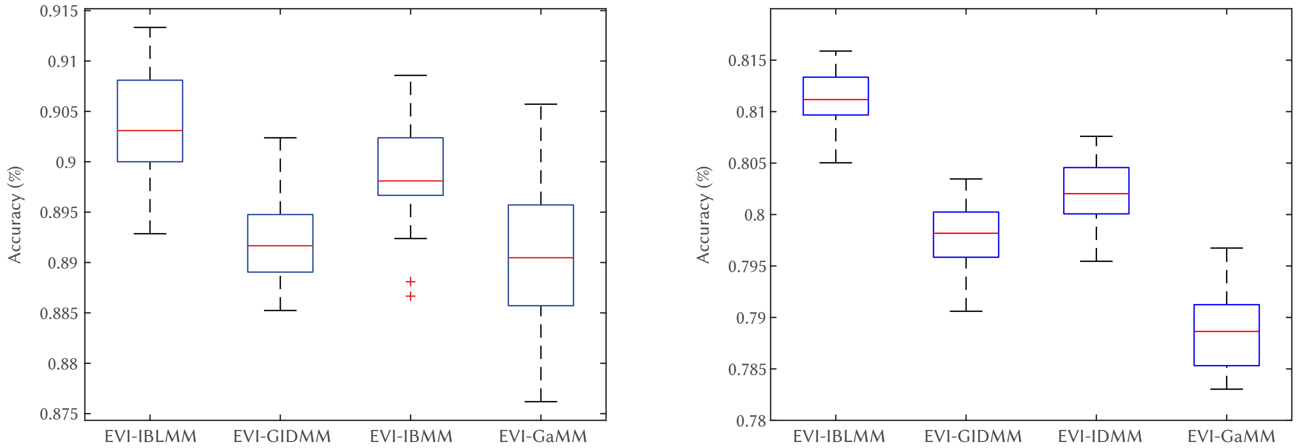
TABLE III. Comparisons of Text Categorization Accuracies (in %) and Runtime (in S) Obtained by Different Approaches

Dataset	Method	EVI-IBLMM	EVI-GIDMM	EVI-IDMM	EVI-GaMM
WebKB	Accuracy	**90.36**	89.27	89.91	89.03
	Runtime	0.66	0.61	0.59	0.39
20Newsgroup	Accuracy	**81.11**	79.82	80.20	78.86
	Runtime	4.85	5.35	3.84	0.71

The text categorization experiment with the proposed EVI-IBLMM in our paper is conducted by using two extensively applied text collections: WebKB [20] and 20Newsgroup[1]. The WebKB dataset is composed of four categories: course, faculty, project and student, with a total of 4,199 documents. The 20Newsgroups dataset contains 13,998 newsgroup documents evenly distributed on 20 categories. Each of these categories is 30 times randomly divided into two separate halves, one half for training and the other half for testing. Following [21], the Porter's stemming [22] is applied to reduce the words to their basic forms. In the pre-processing step, the words that occur less than 3 times or are shorter than 2 in length are eliminated, which results in the representation of each document by a positive vector. The vectors in the different training sets are then modeled by the IBLMM trained by the algorithm in the previous section. Finally, each document vector is categorized to a given category according to the well-known Bayes classification rule.

Three referred methods, namely EVI-based Bayesian GIDMM [13] (EVI-GIDMM), EVI-based Bayesian IDMM (EVI-IDMM) [6] and EVI-based Bayesian Gamma mixture model (EVI-GaMM) [4] are also used to the aforementioned task. Table III shows the mean results of the tested methods in terms of categorization accuracy and training time over 20 runs. Fig. 4 illustrates the categorization accuracies obtained by different methods. Based on these results, it can be found that the proposed EVI-IBLMM has the best categorization accuracy (%) among all the referred mixture-based approaches for the task of text categorization. Moreover, to investigate more insights for the EVI-IBLMM algorithm, the EVI-IBLMM is further compared with deep neural networks (DNNs) on the text categorization task. The fully connected (FC) neural networks with different numbers (i. *e*., *l*) of

1 http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html



(a) WebKB

(b) 20Newsgroup

Fig. 4. Boxplots for comparisons of the categorization accuracies' distributions for the WebKB and the 20Newsgroup datasets.

hidden layers are used. The extracted feature vectors for the WebKB and 20Newsgroup datasets are used as inputs, respectively. These feature vectors are named as shallow feature vectors. The is set as 1,2 , and 4 , respectively and the number of nodes in each hidden layer is the same as the dimension of the shallow features. Table IV shows the comparison of categorization accuracies and training time of different FC neural networks and the proposed EVI-IBLMM algorithm on both WebKB and 20Newsgroup datasets. According to these results, it can be found that the proposed method significantly decreases training time compared to the FC neural networks. Although the proposed approach cannot outperform the DNNs, it can effectively model the features extracted and obtain proper classification accuracies on the two datasets, which can explicitly show the effectiveness of the proposed method.

V. Conclusions

In this paper, an efficient attractive EVI algorithm for the inverted Beta-Liouville mixture model is proposed. Different from the traditional EM algorithm and MCMC algorithm, this algorithm is able to automatically and simultaneously determine all the model's parameters and the optimal number of components, which can prevent the problem of over-fitting. Besides, the proposed algorithm can converge in a short time, and therefore, it has a relatively high efficiency. The good performance of the proposed method is experimentally demonstrated through both synthetic datasets and real datasets which are generated from a real-world application namely text categorization. A future work can be devoted to investigate how to combine a feature selection criterion with the model selection in a unified Bayesian framework or to extend the IBLMM to the infinite case applying some nonparametric Bayesian methods.

Acknowledgment

This work was supported by the General Project of Science and Technology Plan of Beijing Municipal Commission of Education (No. KM201910009014) and the National Natural Science Foundation of China (Grant No. 62172193).

References

[1] M. A. Mashrgy, T. Bdiri, N. Bouguila, "Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models," *Knowledge-Based Systems*, vol. 59, no. 2, pp. 182-195, 2014.

[2] T. Bdiri, N. Bouguila, "Variational bayesian inference for infinite generalized inverted dirichlet mixtures with feature selection and its application to clustering," *Applied Intelligence*, vol. 40, pp. 507-525, 2016.

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

[4] Y. Lai, H. Cao, L. Luo, Y. Zhang, F. Bi, X. Gui, Y. Ping, "Extended variational inference for gamma mixture model in positive vectors modeling," *Neurocomputing*, vol. 432, pp. 145-158, 2021.

[5] T. Bdiri, N. Bouguila, "Positive vectors clustering using inverted dirichlet finite mixture models," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1869-1882, 2012.

[6] Y. Lai, Y. Ping, W. He, B. Wang, J. Wang, X. Zhang, "Variational bayesian inference for finite inverted dirichlet mixture model and its application to object detection," *Chinese Journal of Electronics*, vol. 27, no. 3, pp. 603-610, 2018.

[7] S. Bourouis, M. A. Mashrgy, N. Bouguila, "Bayesian learning of finite generalized inverted dirichlet mixtures: Application to object classification and forgery detection," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2329-2336, 2014.

[8] C. Hu, W. Fan, J. Du, N. Bouguila, "A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models," *Neurocomputing*, vol. 333, pp. 110-123, 2019.

[9] Y. Agusta, D. L. Dowe, "Unsupervised learning of gamma mixture models using minimum message length," in *Proceedings of the 3rd IASTED international multi-conference artificial intelligence and applications*, 2003, pp. 457-462.

[10] S. Favaro, Y. W. Teh, "Mcmc for normalized random measure mixture models," *Statistical Science*, vol. 28, no. 3, pp. 335-359, 2013.

[11] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 2001, pp. 362-369.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

[13] Z. Ma, J. Xie, Y. Lai, J. Taghia, J. Xue, J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-gaussian structured data modeling," *IEEE Transactions on Neural Network and Learning System*, vol. 31, no. 7, pp. 2240-2254, 2019.

[14] W. Fan, N. Bouguila, "Modeling and clustering positive vectors via nonparametric mixture models of liouville distributions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3193-3203, 2020.

[15] G. J. McLachlan, D. Peel, *Finite Mixture Models*. Wiley, New York, USA, 2000.

[16] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, A. Leijon, "Bayesian estimation of dirichlet mixture model with variational inference," *Pattern Recognition*, vol. 47, no. 9, pp. 3143-3157, 2014.

[17] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, O. Deussen, "Document cards: A top trumps visualization for documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1145-1152, 2009.

[18] W. W. Chu, D. B. Johnson, H. Kangarloo, "A medical digital library to support scenario and user-tailored information retrieval," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 2, pp. 97-107, 2000.

[19] Y. Peng, J. Wang, L. Jiao, "A novel text retrieval algorithm for public crisis cases," *Chinese Journal of Electronics*, vol. 28, no. 4, pp. 712-717,

[20] K. Nigam, A. McCallum, S. Thrun, T. Mitchell, "Learning to classify text from labeled and unlabeled documents," in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998, pp. 792-799.

[21] Y. Ping, Y. Zhou, C. Xue, Y. Yang, "Efficient representation of text with multiple perspectives," *The Journal of China Universities of Posts and Telecommunications*, vol. 1, no. 19, pp. 101-111, 2012.

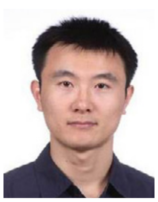
[22] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

Yongfa Ling

Yongfa Ling Ph.D, he currently is the assistant to the President of Hezhou University, Dean of the School of Artificial Intelligence (Modern Industry School), His research interests include Wireless Sensor Network and Communication Network Protocol, Compressed Sensing, MAC Mechanism and System Simulation, Collision Decomposition Algorithm, Priority Control Technology.

Wenbo Guan

Wenbo Guan received the bachelor's degree from the Beijing International Studies University and the master's degree from the North China University of Technology, now he is currently pursuing the Ph.D. degree in the Institute of Information Engineering, CAS. His research interests include pattern recognition, machine learning and cyberspace security.



Qiang Ruan

Ruan Qiang, CEO of DigApis Information Security Technology (Jiangsu) Co., Ltd., engaged in cyberspace security research for more than 20 years. His research interests include network attack and defense, industrial Internet security, and vulnerability mining.

Heping Song

Heping Song received the Ph.D. degree in computer application technology from Sun Yat-sen University, Guangzhou, China, in 2011. He is currently a Associate Professor with the Department of Software Engineering, School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His current research interests include low-level vision, inverse problems, computer vision and deep learning.

Yuping Lai

Yuping Lai from Beijing University of Posts and Telecommunications, Beijing, China, in 2014. He is currently an associate professor in the School of Cyberspace Security at Beijing University of Posts and Telecommunications, Beijing, China. His research interests include information security, machine learning, and data mining.