Extended Variational Inference for A Dirichlet Process Mixture of Beta-Liouville Distributions for Proportional Data Modeling

Yuping Lai, *Member, IEEE,* Lijuan Luo, Yuan Ping, Heping Song, and Hongying Meng, *Senior Member, IEEE*

Abstract—This paper addresses the Bayesian estimation of parameters in the Dirichlet process of the Beta-Liouville distribution (*i.e.*, an infinite Beta-Liouville mixture model (InBLMM)), which has recently gained considerable attentions due to its modeling capability for proportional data. By applying the conventional variational inference (VI) framework, we cannot derive an analytically tractable solution since the variational objective function cannot be explicitly calculated. Therefore, this paper adopts the recently proposed extended VI (EVI) framework to derive a closed-form solution by further lower bounding the original variational objective function in the VI framework. This method is capable of simultaneously determining the model's complexity and estimating the models parameters. Moreover, due to the nature of Bayesian nonparametric approaches, it can also avoid the problems of underfitting and overfitting. Extensive experiments were conducted on both synthetic and real data, generated from two real-world challenging applications namely object detection and text categorization, to evaluate superior performance and effectiveness of the proposed method.

Index Terms—Infinite mixture model, Dirichlet process, Beta-Liouville distribution, extended variational inference, Bayesian estimation, object detection, text categorization.

1 Introduction

Statistical modeling is an important problem in an wide range of research domains, such as artificial intelligence, machine learning, pattern recognition, data mining [1]–[3], and so forth. There exists a myriad of statistical modeling approaches. Among these methods, the finite mixture modeling technique [2] is perhaps the most widely acknowledged and applied statistical modeling approach, which is capable of flexible and powerful probabilistic building tools to describe multimodal distributions of observed data from science, scholarship and daily life. Recently finite mixture models have been applied in a wide variety of real-world applications, such as skin detection [4], speaker identification [5], vehicles detection [6], text categorization [7], and image segmentation [8], [9].

One key and challenging problem in finite mixture modeling is the selection of the probability density functions (PDFs) of components, which depends on the nature of data, and determines the flexibility and robustness of mixture models. Much of early work [5], [10]–[13] adopted the Gaussian distribution as the component density mainly due to its simplicity and maturity. However, the Gaussian distributed assumption is not realistic, since the observations that we would like to model may be non-Gaussian distributed. Such typical non-Gaussian data includes heavy-tailed data [14], [15], skewed data [16], [17], proportional data [18], positive data [19]–[21] and axially symmetric data [22], [23]. Recent research outcomes showed that non-Gaussian mixture models, such as finite Waston mixture model (WMM) [22], finite von Mises-Fisher mixture model (vMFMM) [24], finite Beta mixture model (BMM) [25], finite Dirichlet mixture model (DMM) [4], [26], [27], finite Beta-Liouville mixture model (BLMM) [28], finite generalized Gamma mixture model (GGMM) [21] and finite generalized inverted Dirichlet mixture model (GIDMM) [3], performed better than the finite Gaussian mixture model (GMM) in many applications involving non-Gaussian data. For example, the WMM and the vMFMM have demonstrated their advantages in modeling axially symmetric data. The DMM and the BLMM have been proved to be more efficient in proportional data modeling.

Another important and challenging problem in finite mixture modeling is to select the appropriate number of components (NoC) in the mixture models (*i.e.*, model selection) [2]. If the NoC is not appropriately selected,

- *Y. Lai is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China.*
- *L. Luo is with the School of Business and Management, Shanghai International Studies University, Shanghai, China.*
- *Y. Ping is with the School of Information Engineering, Xuchang University, Xuchang, China.*
- *H. Song is with School of Computer Science and Communications Engineering, Jiangsu University, Zhenjiang, China.*
- *H. Meng is with Electronic and Electrical Engineering Department, Brunel University London, U.K.*
- *Yuping Lai and Lijuan Luo have contributed equally to this work. The corresponding author is H. Song. Email: songhp@ujs.edu.cn, H. Meng. Email: hongying.meng@brunel.ac.uk*

the model tends to underfit or overfit the observed data. Numerous approaches have been developed to tackle this problems, which can be broadly divided into two types, namely deterministic and Bayesian methods. The deterministic methods is pervasively implemented by combining the maximization likelihood estimation (MLE) with penalized log-likelihood criteria such as Bayesian information criterion (BIC) [10] or minimum message length (MML) [29], within the conventional expectation maximization (EM) framework [30]. However, the EM algorithms for non-Gaussian mixture models have three main shortcomings: 1) it is particularly sensitive to the initialization of the model; 2) it is prone to overfitting and converging to local maxima, due to its greedy nature; and 3) the iterative numerical calculation in the maximization step (*e.g.*, with the Newton-Raphson method) incurs prohibitive computational costs. On the other hand, the Bayesian methods allow us to approximate a full Bayesian posterior by means of incorporating prior knowledge about parameters into the models and then marginalizing over parameter uncertainty [31], [32], which avoid the shortcomings related to deterministic techniques. Nevertheless, most of these two types of methods are computationally expensive since they need the evaluation of a given selection criterion for several NoCs, which limits their usages to small-scale problems in practice.

Bayesian non-parametric (BNP) methods [33], [34] are excellent alternatives to the aforementioned methods to deal with the model selection problem in parametric finite mixture modeling, which can be well adapted to depict complex and realistic datasets, and automatically infer the optimal NoC from the data. Among all the BNP methods, the Dirichlet process mixture (DPM) models [23], [35]–[37] attracted the most attentions, which are widely applied to build probability models with flexible latent structures and complexities. A prominent and well-studied example is the infinite mixture modeling, which assumes that the observed data are governed by an infinite NoC, but only a finite NoC does truly generates the data. The majority of early research work with respect to infinite mixture modeling were interested in the infinite Gaussian mixture model (InGMM) [38]. Nevertheless, recent work has shown that DPM with non-Gaussian components, such as infinite Watson mixture model [23] (InWMM), infinite inverted Dirichlet mixture model (InIDMM) [20], infinite Beta-Liouville mixture model (InBLMM) [39], and infinite Beta mixture model (InBMM) [40], are able to provide better modeling ability than the InGMM in the case of non-Gaussian distributed data.

An essential problem in the BNP modeling is to infer the posterior distributions of the latent variables. However, it is infeasible to evaluate the posterior distribution, due to the coupling among these variables. To address this problem, practitioners need to resort to approximation inference methods, which can be broadly divided into types, namely stochastic and deterministic techniques. The stochastic techniques are mathematical solutions that rely on repeated random numerical sampling to obtain their results. Markov chain Monte Carlo (MCMC) [41] approaches are the most extensively applied stochastic technique. Nonetheless, the MCMC methods are prone to suffer from bad performance, as two shortcomings exist: 1) it is difficult to diagnose their convergences, especially when working with large-scale data; and 2) the sampling procedure is time-consuming and computationally expensive, such that they cannot be extended to fit large datasets and high model complexities. On the other hand, the deterministic techniques applied a computationally tractable parametric distribution to approximate the actual posterior. The VI is the most widespread deterministic method as a result of its excellent generalization performance and computational simplicity in a variety of applications, including finite mixture models learning [10], [20]–[22], [24], [26], which can provide an excellent alternative to the MCMC-based sampling algorithms. The main idea of the VI is to utilize a simple family of distributions to approximate the analytically intractable true posterior and then discover a member from this family which is as close as the actual posterior. The Kullback-Leibler (KL) divergence from the approximated posterior to the actual posterior is applied to measure the closeness. Minimizing the KL divergence is then turned into an optimization problem of the evidence likelihood bound (ELBO) [3], [42].

Motivated by the powerful modeling capabilities of the DPM and the excellent performance of the VI framework, this work focuses on Bayesian estimation of the DPM of the Beta-Liouville distributions (*i.e.*, InBLMM) with the VI framework. Unfortunately, applying the traditional VI framework cannot derive an analytically tractable solution for the Bayesian estimation of the InBLMM, since the calculation of the ELBO requires intractable moment computations. The EVI framework [20], [27] is a good alternative to the VI framework, which has proved to be efficient in non-Gaussian mixtures learning [20], [25], [27], [43], [44]. Following the principles of the EVI, lower-bound approximations, which are subject to certain constraints [27], are introduced to the ELBO in the traditional VI framework. With these auxiliary functions-based lower-bound approximations in hand, we can easily derive a closed-form solution for the InBLMM. Our major contributions of this article are summarized into three aspects. First, the BLMM has been extended to the InBLMM via the stick-breaking representation, which provides an elegant way to simultaneously carry out parameter estimation and model selection. Second, a closed-form solution for Bayesian estimation of the InBLMM is derived through employing the EVI framework. This method is capable of guaranteeing theoretical convergence and providing better approximations. Third, the proposed method has been applied for object detection and text categorization. The experimental results upon both artificial and realistic datasets demonstrate the better performance of the proposed method than the referred methods.

This paper is organized as follows. We first briefly review the BLMM and the DPM model, and extend the BLMM into the infinite case in Section 2. Then, we detail a complete EVI framework for learning the InBLMM in Section 3. The experimental results and comparisons are given in Section 4. Finally, we draw some conclusions and future works in Section 5

2 Model Specification

2.1 Finite Beta-Liouville Mixture Model (BLMM)

Let $\mathbf{x} = [\mathrm{x}_1, \cdots, \mathrm{x}_D]^{\mathrm{T}}$ be a D-dimensional random proportional vector, where $0 < \mathrm{x}_d < 1$ for $d = 1, \cdots, D$, and $\sum_{d=1}^{D} \mathrm{x}_d < 1$. We can apply a Beta-Liouville (BL) distribution to model its underlying distribution. Several researches have validated the advantages of choosing the BL distribution in modeling proportional vectors [39], [45], [46]. The probability density function (PDF) of a BL distribution is given by [47]

$$\begin{aligned}\mathrm{BL}(\mathbf{x}|\boldsymbol{\theta}) =& \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \prod_{d=1}^{D} \frac{\mathrm{x}_{\mathrm{d}}^{\alpha_{\mathrm{d}}\text{-}1}}{\Gamma(\alpha_d)} \\ &\times \left(\sum_{d=1}^{D} \mathrm{x}_d\right)^{u-\sum_{d=1}^{D}\alpha_d} \left(1-\sum_{d=1}^{D} \mathrm{x}_d\right)^{v-1},\end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = [\alpha_1, \cdots, \alpha_D, u, v]^{\mathrm{T}}$ is the parameter vector that consists of only positive real values, and $\Gamma(\cdot)$ denotes the Gamma function. Note that the Dirichlet distribution [26], [27] is a special case of the BL distribution. To flexibly model the proportional data with multimodal distributions, a mixture modeling technique [48] is used to build a BLMM. With M mixture components, the PDF of the BLMM is defined as [18]

$$p(\mathbf{x}|\boldsymbol{\Pi}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m \mathrm{BL}(\mathbf{x}|\boldsymbol{\theta}_m), \tag{2}$$

where $\boldsymbol{\Pi} = [\pi_1, \cdots, \pi_M]^{\mathrm{T}}$ are the mixing proportions and satisfy the constraints as

$$\pi_m \geq 0, \quad \text{and} \quad \sum_{m=1}^{M} \pi_m = 1. \tag{3}$$

In addition, $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M]$ is the parameter matric. For proportional data from many real-world applications, the BLMM, among others, has been extensively used for the purpose of clustering and classification of such data [18], [28].

2.2 Infinite BL Mixture Model with Stick Breaking

The Dirichlet process (DP) [49] is a famous stochastic process that is extensively applied in BNP models of data, particularly in DPM models. To avoid problems related to model selection, we extend the BLMM to a case with infinite NoC through applying the framework of the DPM model, and construct the DPM model of the BL distributions (which we refer to as the InBLMM) via the following stick-breaking representation [50]–[52]. Let G be DP distributed as $G \sim DP(\phi, H)$, where ϕ is a positive scaling parameter and H is a base distribution. The stick-breaking representation of G is as

$$\begin{aligned} G &= \sum_{m=1}^{\infty} \pi_m \delta_{\Omega_m}, \\ \pi_m &= \lambda_m \prod_{j=1}^{m-1} (1-\lambda_j), \lambda_m \sim \mathrm{Beta}(1, \phi), \\ \Omega_m &\sim H, \end{aligned} \tag{4}$$

where δ_{Ω_m} denotes the Dirac measure with unit mass at Ω_m. π_m are the non-negative mixing weights that sum to one and can imaginarily be given by breaking a stick of unit length into a countably infinite number of pieces.

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be a sample of N observations, which are independently drawn from an InBLMM. For each observation $\mathbf{x}_n$, we have a corresponding hidden component indicator vector variable $\mathbf{z}_n = \{\mathrm{z}_{n1}, \mathrm{z}_{n2}, \cdots\}$, which is defined as $\mathrm{z}_{nm} \in \{0, 1\}$, $\sum_{m=1}^{\infty} \mathrm{z}_{nm} = 1$, and $\mathrm{z}_{nm} = 1$ given that $\mathbf{x}_n$ comes from the mth component and 0 otherwise. Therefore, the conditional distribution of dataset $\mathbf{X}$ given the hidden variables $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_N]$ is given by

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \mathrm{BL}(\mathbf{x}_n|\boldsymbol{\theta}_m)^{\mathrm{z}_{nm}}. \tag{5}$$

The prior distribution of the latent variables $\mathbf{Z}$ given the mixing weights $\boldsymbol{\Pi}$ is given by

$$p(\mathbf{Z}|\boldsymbol{\Pi}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \pi_m^{\mathrm{z}_{nm}}. \tag{6}$$

Because $\boldsymbol{\Pi}$ is a function of $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \cdots\}$ as shown in (4), we can rewrite $p(\mathbf{Z})$ as

$$p(\mathbf{Z}|\boldsymbol{\lambda}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \left[\lambda_m \prod_{j=1}^{m-1} (1-\lambda_m)\right]^{\mathrm{z}_{nm}}. \tag{7}$$

The variable $\boldsymbol{\lambda}$ in (4) has a specific Beta distribution as

$$p(\boldsymbol{\lambda}|\boldsymbol{\eta}) = \prod_{m=1}^{\infty} \mathrm{Beta}(\lambda_m|1, \eta_m) = \prod_{m=1}^{\infty} \eta_m(1-\lambda_m)^{\eta_m - 1}, \tag{8}$$

where $\boldsymbol{\eta} = \{\eta_1, \eta_2, \cdots\}$ are hyperparameters.

To complete the Bayesian formulation of the InBLMM, proper conjugate priors need to be imposed over the parameters $\boldsymbol{\theta}_m$ of the BL distributions. Motivated by [25], we define these parameters to be Gamma distributed as

$$p(\boldsymbol{\Lambda}) = \mathcal{G}(\boldsymbol{\Lambda}; \mathbf{S}, \mathbf{T}) = \prod_{m=1}^{\infty} \prod_{d=1}^{D} \frac{t_{md}^{s_{md}}}{\Gamma(s_{md})} e^{-s_{md}\alpha_{md}}, \tag{9}$$

$$p(\mathbf{U}) = \mathcal{G}(\mathbf{U}; \mathbf{G}, \mathbf{H}) = \prod_{m=1}^{\infty} \frac{h_m^{g_m}}{\Gamma(g_m)} e^{-g_m u_m}, \tag{10}$$

$$p(\mathbf{V}) = \mathcal{G}(\mathbf{V}; \mathbf{P}, \mathbf{Q}) = \prod_{m=1}^{\infty} \frac{q_m^{p_m}}{\Gamma(p_m)} e^{-p_m v_m}, \tag{11}$$

where the hyperparameters $\mathbf{\Lambda} = \{\alpha_{md}\}$, $\mathbf{U} = \{u_m\}$, and $\mathbf{V} = \{v_m\}$ denote the random variable sets. Moreover, the hyperparameters $\mathbf{S} = \{s_{md}\}$, $\mathbf{T} = \{t_{md}\}$, $\mathbf{G} = \{g_m\}$, $\mathbf{H} = \{h_m\}$, $\mathbf{P} = \{p_m\}$, and $\mathbf{Q} = \{q_m\}$ are strictly positive. The joint density function of all random variables in the InBLMM is then given by Fig. 1.

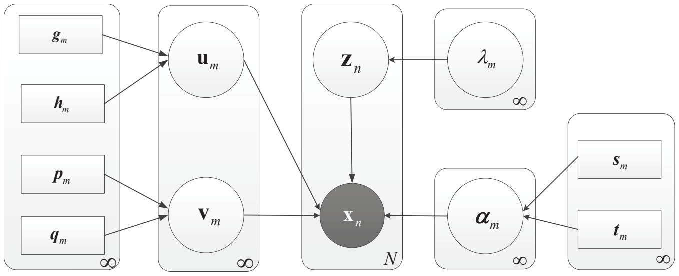


Fig. 1: Graphical model representation of the InBLMM. Circles represent random variables, boxes represent hyperparameters, and the arrows show the conditional dependence between variables.

The joint distribution of all the random variables can be obtained by applying Bayes' theorem and combining (5) and (6)-(11) as

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{\Theta}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}|\boldsymbol{\eta}) p(\mathbf{\Lambda}) p(\mathbf{U}) p(\mathbf{V}) \\
&= \prod_{n=1}^{N} \prod_{m=1}^{\infty} \Bigg[\lambda_m \prod_{j=1}^{m-1} (1 - \lambda_m) \frac{\Gamma(\sum_{d=1}^{D} \alpha_{md})}{\prod_{d=1}^{D} \Gamma(\alpha_{md})} \\
&\times \frac{\Gamma(u_m + v_m)}{\Gamma(u_m)\Gamma(v_m)} \prod_{d=1}^{D} \mathrm{x}_{nd}^{\alpha_{md}\text{-}1} \left(1 - \sum_{d=1}^{D} \mathrm{x}_{nd}\right)^{v_m - 1} \\
&\times \left(\sum_{d=1}^{D} \mathrm{x}_{nd}\right)^{u_m - \sum_{d=1}^{D} \alpha_{md}} \Bigg]^{\mathrm{z}_{nm}} \prod_{m=1}^{\infty} \prod_{d=1}^{D} \frac{t_{md}^{s_{md}}}{\Gamma(s_{md})} \\
&\times \alpha_{md}^{s_{md} - 1} e^{-t_{md} \alpha_{md}} \prod_{m=1}^{\infty} \Bigg[\eta_m (1 - \lambda_m)^{\eta_m - 1} \frac{h_m^{g_m}}{\Gamma(g_m)} \\
&\times u_m^{g_m - 1} e^{-h_m u_m} \frac{q_m^{p_m}}{\Gamma(p_m)} v_m^{p_m - 1} e^{-q_m v_m} \Bigg],
\end{aligned} \tag{12}$$

where we have defined $\mathbf{\Theta} = \{\mathbf{Z}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{V}\}$ to simplify the notations. In the next section, we will propose a novel variational approximation scheme for InBLMM, which can simultaneously handle the issue of estimating model parameters and choosing the optimal number of mixture components.

3 Model Learning

In this section, we develop a Bayesian estimation approach for the InBLMM based upon the EVI framework [20], [44], [53], which is a variant of the conventional VI framework [3]. The VI treatment of the InBLMM is conducted through introducing an arbitrary distribution $q(\mathbf{\Theta}) = q(\mathbf{Z}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{V})$ to approximate the actual posterior distribution $p(\mathbf{\Theta}|\mathbf{X})$, and considering the well-known equality for the log-marginal likelihood (log evidence) $\ln p(\mathbf{X})$ as

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p), \tag{13}$$

where we have defined

$$\mathcal{L}(q) = -\int_{\mathbf{\Theta}} q(\mathbf{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}. \tag{14}$$

In (14), $\mathrm{KL}(q||p)$ stands for the Kullback-Leibler (KL) divergence between the approximated posterior $q(\mathbf{\Theta})$ and the actual posterior $q(\mathbf{X}|\mathbf{\Theta})$, which is given by

$$\mathbf{KL}(q||p) = -\int_{\mathbf{\Theta}} q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta}|\mathbf{X})}{q(\mathbf{\Theta})} d\mathbf{\Theta}. \tag{15}$$

Because the KL divergence is nonnegative and is zero when $q(\mathbf{\Theta}) = p(\mathbf{\Theta}|\mathbf{X})$, $\mathcal{L}(q)$ is a rigorous lower bound of the log evidence, *i.e.*, $\ln p(\mathbf{X}) \geq \mathcal{L}(q)$. Minimization of the KL divergence is equivalent to maximization of the lower bound. Nevertheless, it is infeasible to solve $q(\mathbf{\Theta})$ through minimizing $\mathrm{KL}(q||p)$, because $q(\mathbf{X}|\mathbf{\Theta})$ is unknown. Hence, optimization of the lower bound is extensively applied in the conventional VI framework to reach an excellent approximation distribution. The lower bound $\mathcal{L}(q)$ in (14) is also known as the variational object function and can be further rewritten as

$$\mathcal{L}(q) = \mathrm{E}_{\mathbf{\Theta}}[\ln p(\mathbf{X}, \mathbf{\Theta})] - \mathrm{E}_{\mathbf{\Theta}}[\ln q(\mathbf{\Theta})]. \tag{16}$$

Unfortunately, for most of the non-Gaussian infinite mixture models, such as the InDMM, the InBMM, and the InBLMM, the object function $\mathcal{L}(q)$ is untractable to compute in a closed-form expression, since the evaluation of $\mathcal{L}(q)$ needs intractable moment computations in $\mathrm{E}_{\mathbf{\Theta}}[\ln p(\mathbf{X}, \mathbf{\Theta})]$. Consequently, directly maximizing $\mathcal{L}(q)$ to solve $q(\mathbf{\Theta})$ is unfeasible and then the VI framework cannot work on non-Gaussian mixtures. The EVI framework [20], [44] can be applied to tackle this issue elegantly. The main idea behind the EVI method is that, if a help function $\tilde{p}(\mathbf{X}, \mathbf{\Theta})$ that satisfies

$$\mathbf{E}_{\Theta}[p(\mathbf{X}, \mathbf{\Theta})] \geq \mathbf{E}_{\mathbf{\Theta}}[\tilde{p}(\mathbf{X}, \mathbf{\Theta})] \tag{17}$$

can be found, then the maximum value of $\mathcal{L}(q)$ can be asymptotically reached by means of maximizing the lower-bound of $\mathcal{L}(q)$ rather than maximizing $\mathcal{L}(q)$ itself. With the aid of (17), we obtain a lower bound of $\mathcal{L}(q)$ as

$$\tilde{\mathcal{L}}(q) = \mathrm{E}_{\mathbf{\Theta}}[\ln \tilde{p}(\mathbf{X}, \mathbf{\Theta})] - \mathrm{E}_{\mathbf{\Theta}}[\ln q(\mathbf{\Theta})]. \tag{18}$$

To obtain a tractable expression for $\tilde{\mathcal{L}}(q)$, we further adopt the mean-field approximation supposition [3], [10] that the variational distribution $q(\mathbf{\Theta})$ can be factorized over the latent variables in $\mathbf{\Theta}$. Moreover, a truncated stick-breaking construction technique [52] is adopted to truncate $q(\mathbf{\Theta})$ through fixing a value M and letting $q(\lambda_M = 1) = 1$, which implies $q(z_{nm})$=0 for $m > M$. Consequently, the variational distribution $q(\mathbf{\Theta})$ can be written as

$$q(\mathbf{\Theta}) = \prod_{n=1}^{N} \prod_{m=1}^{M} q(\mathrm{z}_{nm}) \prod_{m=1}^{M} \prod_{d=1}^{D} q(\alpha_{md}) q(\beta_{md}) \prod_{m=1}^{M} q(\lambda_m). \tag{19}$$

Note that the hidden variables in $\mathbf{\Theta}$ generally have different variational parameters and no restriction is placed upon individual variational factors [3].

With the assumption of the factorization formulation, the optimal solution to each variational factor can be obtained via maximization of $\tilde{\mathcal{L}}(q)$ in (19). The optimal solution for a specified factor $q_j(\Theta_j)$ within the EVI framework is then given by

$$\ln q_j(\Theta_j) = \langle \ln \tilde{p}(\mathbf{X}, \mathbf{\Theta}) \rangle_{s \neq j} + \text{Con.}, \tag{20}$$

where $\langle \cdot \rangle_{s \neq j}$ indicates an expectation regarding all the factors $q_s(\Theta_s)$ except for $s = j$. "Con." denotes a constant that is independent of Θ_j and is employed to normalize the corresponding factor. Note that EVI represents a factor using knowledge about other factors; therefore, it is essentially iterative.

Applying (19) yields optimal variational posteriors as (proofs are shown in Appendix A):

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{m=1}^{M} r_{nm}^{\mathrm{z}_{nm}}, \tag{21}$$

$$q(\boldsymbol{\lambda}) = \prod_{m=1}^{M} \text{Beta}(\lambda_m | a_m^*, b_m^*), \tag{22}$$

$$q(\mathbf{U}) = \prod_{m=1}^{M} \mathcal{G}(u_m | g_m^*, h_m^*), \tag{23}$$

$$q(\mathbf{V}) = \prod_{m=1}^{M} \mathcal{G}(v_m | p_m^*, q_m^*), \tag{24}$$

$$q(\mathbf{\Lambda}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{G}(\alpha_{md} | s_{md}^*, t_{md}^*). \tag{25}$$

In (21)-(25), the hyperparameters are obtained through maximization and determination of the density involved in $q(\mathbf{\Theta})$. The estimates of r_{nm} in (21) is given by

$$r_{nm} = \frac{\rho_{nm}}{\sum_{n=1}^{N} \rho_{nm}}, \tag{26}$$

where

$$\begin{aligned} \ln \rho_{nm} =& \tilde{\mathcal{R}}_m + \tilde{\mathcal{F}}_m + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1 - \lambda_j) \rangle \\ & + \sum_{d=1}^{D} (\bar{\alpha}_{md} - 1) \ln \mathrm{x}_{nd} + (\bar{v}_m - 1) \ln(1 - \sum_{d=1}^{D} \mathrm{x}_{nd}) \\ & + (\bar{u}_m - \sum_{d=1}^{D} \bar{\alpha}_{md}) \ln(\sum_{d=1}^{D} \mathrm{x}_{nd}). \end{aligned} \tag{27}$$

In (27), $\tilde{\mathcal{R}}_m$ and $\tilde{\mathcal{F}}_m$ are defined in (38) and (39), respectively. The estimates of a_m^* and b_m^* at the new steps are given by

$$a_m^* = 1 + \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle, b_m^* = \eta_m + \sum_{n=1}^{N} \sum_{j=m+1}^{M} \langle \mathrm{z}_{nj} \rangle. \tag{28}$$

The estimates of s_{md}^* and t_{md}^* are given by

$$s_{md}^* = s_{md} + \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \left[\Psi(\sum_{d=1}^{D} \bar{\alpha}_{md}) - \Psi(\bar{\alpha}_{md}) \right] \bar{\alpha}_{md}, \tag{29}$$

$$t_{md}^* = t_{md} - \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \left[\ln \mathrm{x}_{nd} - \ln(\sum_{d=1}^{D} \ln \mathrm{x}_{nd}) \right]. \tag{30}$$

The estimates of g_m^* and h_m^* are given by

$$g_m^* = g_m + \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \left[\Psi(\bar{u}_m + \bar{v}_m) - \Psi(\bar{u}_m) \right] \bar{u}_m, \tag{31}$$

$$h_m^* = h_m - \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \ln(\sum_{d=1}^{D} \mathrm{x}_{nd}). \tag{32}$$

The estimates of p_m^* and q_m^* are given by

$$p_m^* = p_m + \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \left[\Psi(\bar{u}_m + \bar{v}_m) - \Psi(\bar{v}_m) \right] \bar{v}_m, \tag{33}$$

$$q_m^* = q_m - \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle \ln(1 - \sum_{d=1}^{D} \mathrm{x}_{nd}). \tag{34}$$

The proofs of the expected values $\langle \cdot \rangle$ included in the above formulas are provided in Appendix B.

From the equations for the variational factors in (21) to (25), the lower bound $\tilde{\mathcal{L}}(q)$ can be evaluated as

$$\begin{aligned} \tilde{\mathcal{L}}(q) =& \langle \ln \tilde{p}(\mathbf{X}, \mathbf{\Theta}) \rangle - \langle \ln q(\mathbf{Z}) \rangle - \langle \ln q(\boldsymbol{\lambda}) \rangle \\ & - \langle \ln q(\mathbf{\Lambda}) \rangle - \langle \ln q(\mathbf{U}) \rangle - \langle \ln q(\mathbf{V}) \rangle, \end{aligned} \tag{35}$$

where the analytical expressions of the terms $\langle \cdot \rangle$ are evaluated regarding all the variables in its argument and are provided in the Appendix C. A complete outline of the developed algorithm of Bayesian estimation of the InBLMM with the EVI is presented in Algorithm 1, and its convergence can be assessed through checking the lower bound $\tilde{\mathcal{L}}(q)$ in (35).

Algorithm 1 Bayesian estimation of the InBLMM with the EVI.

```
Set the initial truncation level $M$.
Initiate the prior distribution parameters $s_{md}$, $t_{md}$, $g_m$, $h_m$, $p_m$, $q_m$ and $\eta_m$.
Initialize $r_{nm}$ via $K$-means algorithm.
repeat
    The variational E-step: Update the expectations in (51)-(52).
    The variational M-step: Update the variational posteriors via (21)-(25).
until Stop criteria are reached.
Calculate $\langle \lambda_m \rangle = a_m^*/(a_m^* + b_m^*)$ and substitute them back into (4) to calculate $\pi_m$.
Detect the optimal $M$ via discarding the components which have very small weights ($\leq 10^{-5}$).
```

4 Experimental Results

Extensive experiments were conducted to assess the effectiveness of the Bayesian estimation method derived in Section 3 on simulated along with real datasets. The simulated data validation aims at evaluating the performance of the Bayesian InBLMM with single lower bound (SLB) approximation (proposed in this paper and referred to as InBLMM$_{\text{SLB}}$) and comparing it with the Bayesian InBLMM with multiple lower bound (MLB) approximation [39] (namely InBLMM$_{\text{MLB}}$). For the details of these two approximation technologies, please refer to [20], [27], [44]. The validations with realistic data are based on two real-life challenging applications, *i.e.*, object detection and text categorization. The aim of real data evaluation is to compare InBLMM$_{\text{SLB}}$ with some recently proposed statistical models, which are all built for modeling proportional vectors. These models include the above InBLMM$_{\text{MLB}}$, the Bayesian InBMM using the SLB approximation (InBMM$_{\text{SLB}}$) [54], the Bayesian InBMM using the MLB approximation (InBMM$_{\text{MLB}}$) [52], the Bayesian InDMM using the SLB approximation (InDMM$_{\text{SLB}}$) [28], [55], and the Bayesian InDMM using the MLB approximation (InDMM$_{\text{MLB}}$) [28]. Moreover, the variational InGMM and SVM were also evaluated and compared with the proposed InBLMM$_{\text{SLB}}$.

At the initialization phase, the truncation level M and the hyperparameters η of the Beta prior are set as equal to $M = 15$ and $\eta = 1$, respectively. Moreover, the hyperparameters s_{md}, g_m, p_m, η_m, t_{md}, h_m, q_m of the Gamma priors are set as $s_{md} = g_m = p_m = \eta_m = 1$ and $t_{md} = h_m = q_m = 0.1$. The prior distributions are thus noninformative. Note that these specific choices were based upon our experimental experiences and also found convenient in our experiments.

4.1 Simulated Data Validation

Firstly, the performance of the proposed InBLMM$_{\text{SLB}}$ on learning the InBLMM is evaluated upon simulated datasets, which were drawn from four known BLMMs. Table 1 and 2 shows the actual parameters for the four BLMMs, and the mean estimated parameters of each dataset applying the InBLMM$_{\text{SLB}}$ and the InBLMM$_{\text{MLB}}$ for each dataset over 20 runs of simulation, respectively. There is obvious evidence that both learning methods are capable to correctly estimate the parameters of the InBLMM, but the InBLMS$_{\text{SLB}}$ always gives more correct results. There exists two methods to estimate the optimal NoC [55]. The first method applies the ELBO as a model selection score, *i.e.*, the variational optimization is implemented upon a fixed truncation level M without updating the mixing proportions and the correct NoC is then selected when the ELBO reaches its maximum value. To validate this approach, we run InBLMS$_{\text{SLB}}$ with M ranging from 1 to 15, and the results are reported in Fig. 2. It is obvious that the ELBO invariantly reach its maximum value at the correct NoC, indicating that discovering the ELOB's global maximum can yield the correct NoC. The second method determines the correct NoC through eliminating the components with tiny mixing proportions after convergence. The estimated mixing proportions of each components for each simulated datasets after convergence are illustrated in Fig. 3. It can be observed that the values of mixing proportions of some components approach zero, which indicates that they are redundant and can be removed. To trace the optimization process of the InBLMS$_{\text{SLB}}$, the value of the ELBO during iterative process is illustrated in Fig. 4. We can clearly see that the ELBO increases by small amounts at each iteration when all mixing proportions are large, and rather quickly when one proportion approaches zero.

TABLE 1: Parameters for generating four simulated datasets: D1 ($N = 500$), D2 ($N = 1000$), D3 ($N = 1000$), D4 ($N = 1000$). N_m denotes the number of elements in cluster m.

Dataset	m	N_m	α_{m1}	α_{m2}	α_{m3}	u_m	v_m	π_m
D1	1	300	24.00	8.00	12.00	24.00	4.00	0.600
	2	200	8.00	12.00	5.00	4.00	8.00	0.400
D2	1	200	14.00	6.00	18.00	20.00	14.00	0.200
	2	300	18.00	21.00	15.00	14.00	28.00	0.300
	3	500	15.00	32.00	10.00	14.00	8.00	0.500
D3	1	150	2.00	6.00	24.00	12.00	28.00	0.150
	2	200	8.00	36.00	15.00	4.00	18.00	0.200
	3	300	48.00	18.00	14.00	18.00	28.00	0.300
	4	350	18.00	24.00	12.00	16.00	8.00	0.350
D4	1	150	12.00	16.00	44.00	32.00	16.00	0.150
	2	200	32.00	48.00	12.00	18.00	12.00	0.200
	3	250	24.00	8.00	34.00	6.00	18.00	0.250
	4	300	12.00	60.00	16.00	25.00	18.00	0.300
	5	100	28.00	12.00	6.00	24.00	8.00	0.100

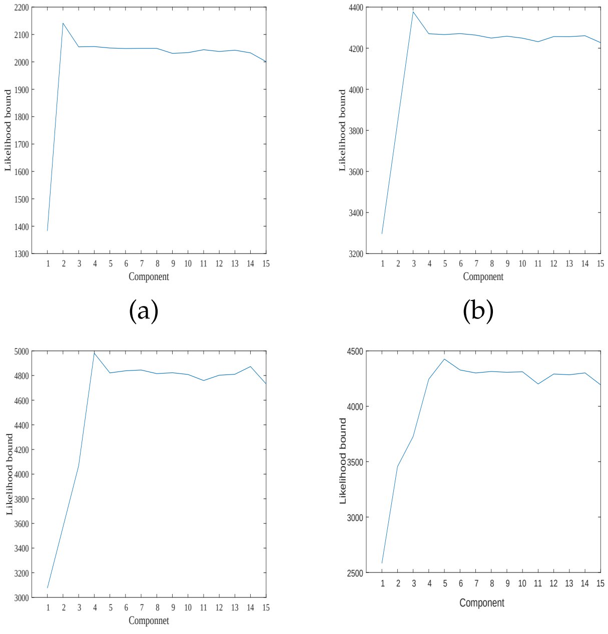
Secondly, we further compare the InBLMM$_{\text{SLB}}$ with the InBLMM$_{\text{MLB}}$ for each simulated dataset to find the better approximation in terms of the ELBO, the KL divergence [1], the convergence time, and the number of iterations before convergence. From these comparisons in Table 3, we can note the following observations: 1) the mean values of the ELBO obtained by the InBLMM$_{\text{SLB}}$ are larger than those obtained by the InBLMM$_{\text{SLB}}$, indicating that the InBLMM$_{\text{SLB}}$ is tighter than the InBLMM$_{\text{MLB}}$; 2) the InBLMM$_{\text{SLB}}$ converges in fewer iterations and takes shorter computational time than the InBLMM$_{\text{MLB}}$; and 3) InBLMM$_{\text{SLB}}$ can yield smaller KL divergence than the InBLMM$_{\text{MLB}}$ for each simulated dataset, indicating that the InBLMM$_{\text{SLB}}$ is capable of discovering a more excellent approximation than the InBLMM$_{\text{MLB}}$ does. Moreover, for the goal of further stability checks, we draw box plots for the distributions of the ELBO and the runtime. Comparisons between the two methods are shown in Fig. 5 and 6. It can be clearly observed that the InBLMM$_{\text{SLB}}$ has more compact lower-bound

[1] Here, we calculate the KL divergence $\text{KL}(p(\mathbf{X}|\mathbf{\Theta})||p(\mathbf{X}|\tilde{\mathbf{\Theta}}))$ via a sampling approach. $\tilde{\mathbf{\Theta}}$ denotes the parameter estimates in the InBLMM.

TABLE 2: Mean estimated parameters for the simulated datasets over 20 runs of $\text{InBLMM}_{\text{SLB}}$ and $\text{InBLMM}_{\text{MLB}}$.

			$\text{InBLMM}_{\text{SLB}}$						$\text{InBLMM}_{\text{MLB}}$					
Dataset	m	N_m	$\hat{\alpha}_{m1}$	$\hat{\alpha}_{m2}$	$\hat{\alpha}_{m3}$	$\hat{u}_m$	$\hat{v}_m$	$\hat{u}_m$	$\tilde{\alpha}_{m1}$	$\tilde{\alpha}_{m2}$	$\tilde{\alpha}_{m3}$	$\tilde{u}_m$	$\tilde{v}_m$	$\tilde{\pi}_m$
D1	1	200	8.15	12.37	5.06	4.01	8.09	0.400	8.24	12.47	5.09	4.05	8.11	0.401
	2	300	23.53	7.78	11.84	24.21	4.04	0.600	23.36	7.79	11.78	24.42	4.07	0.599
D2	1	200	14.02	6.04	17.99	20.21	14.03	0.200	13.91	5.89	17.88	21.12	13.91	0.200
	2	300	18.08	21.21	15.07	13.72	27.89	0.300	18.17	21.34	15.13	13.79	27.39	0.300
	3	500	15.09	32.08	9.98	14.31	8.14	0.500	15.12	32.18	9.98	14.39	8.18	0.500
D3	1	150	1.99	6.01	23.91	11.75	27.42	0.15	1.97	6.06	24.14	11.69	27.29	0.150
	2	200	8.08	36.43	15.17	4.01	18.05	0.200	8.21	36.88	15.26	3.98	17.89	0.200
	3	300	47.26	17.69	13.77	17.95	28.09	0.300	46.75	17.49	13.62	17.66	27.57	0.300
	4	350	17.81	23.73	11.88	16.08	8.03	0.350	17.74	23.71	11.81	16.29	8.37	0.350
D4	1	150	11.74	15.78	42.74	30.58	15.24	0.150	11.63	15.64	42.37	30.52	15.21	0.150
	2	200	29.99	46.18	11.73	18.00	12.00	0.200	29.92	45.33	11.36	18.17	12.16	0.200
	3	250	23.42	7.81	33.18	6.11	18.41	0.250	23.35	7.78	33.05	6.18	18.61	0.2500
	4	300	11.82	58.91	15.73	25.39	18.17	0.300	11.65	58.13	15.52	25.26	18.18	0.300
	5	100	28.13	12.18	6.16	23.61	7.78	0.100	27.86	11.78	6.24	23.42	7.64	0.100

and runtime range and larger median. These results illustrate that the $\text{InBLMM}_{\text{SLB}}$ performs better than the $\text{InBLMM}_{\text{MLB}}$.



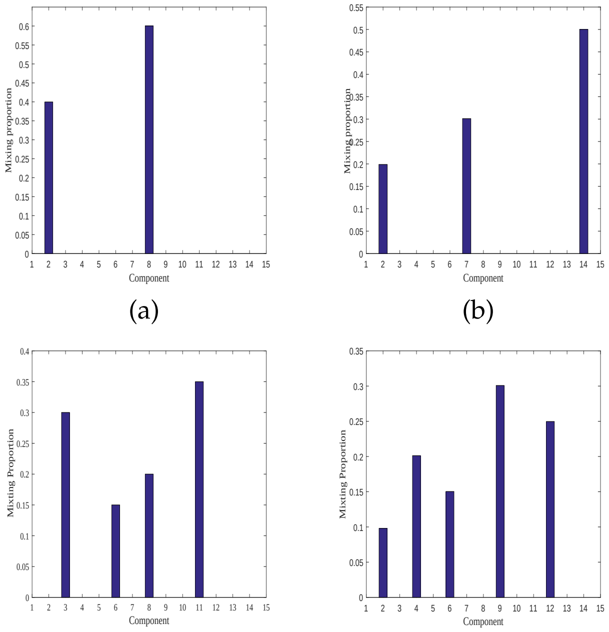
(a)

(b)

(c)

(d)

Fig. 2: ELBO as a function of the fixed supposed NoC for different simulated datasets. (a) D1. (b) D2. (c) D3. (d) D4.



(a)

(b)

(c)

(d)

Fig. 3: Mixing proportions of mixture components found for each simulated datasets after convergence. (a) D1. (b) D2. (c) D3. (d) D4.

4.2 Real Data Evaluation

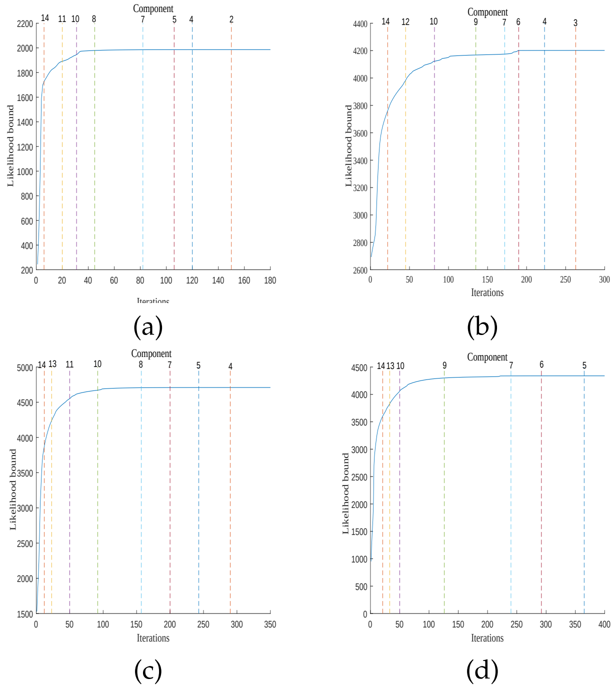
4.2.1 *Object Detection*

Object detection refers to the process of discovering instances of objects from the given categories (such as cars, faces, motorbikes, and airplanes) in some given images or videos. A lot of computer vision (CV) researches [56], [57] have focused on object detection during the few past years, since it has widespread its applications including intelligent traffic management [58], video surveillance [59], medical image analysis [60], and human-computer interfaces [61]. Since pose and illumination changes, scale variations, and occlusions and intra-class variability are prone to yield highly variable images, this task is still difficult and critical in the CV. There exists two categories of insightful approaches to tackling these issues. One concentrates on developing excellent image descriptors [62], [63], and the other one focuses on developing powerful and robust classifiers [64], [65].

A key step for achieving high object detection accuracy is to extract robust descriptors which can be applied for the effective representation of these images. Recently, CV researchers have proposed a variety of valuable global and local visual descriptors. Among all the descriptors,

TABLE 3: The mean ELBO, KL divergence, convergence time, and number of iterations (NoI.) for different simulated datasets of InBLMM$_{\mathrm{SLB}}$ and InBLMM$_{\mathrm{MLB}}$.

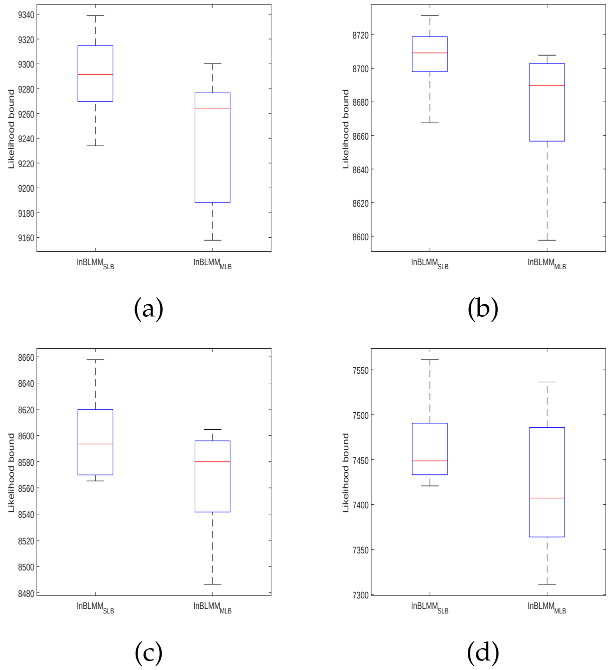
	D1		D2		D3		D4	
	InBLMM$_{\mathrm{SLB}}$	InBLMM$_{\mathrm{MLB}}$	InBLMM$_{\mathrm{S}}$	InBLMM$_{\mathrm{MLB}}$	InBLMM$_{\mathrm{SLB}}$	InBLMM$_{\mathrm{SLB}}$	InBLMM$_{\mathrm{SLB}}$	InBLMM$_{\mathrm{SLB}}$
ELBO	2.009×10^3	1.984×10^3	4.231×10^3	4.209×10^3	4.796×10^3	4.776×10^3	4.358×10^3	4.337×10^3
p-values	7.11×10^{-3}		0.197		0.112		2.35×10^{-4}	
KL	4.12×10^{-5}	7.59×10^{-5}	2.66×10^{-4}	3.95×10^{-4}	6.17×10^{-6}	9.01×10^{-6}	1.06×10^{-4}	4.08×10^{-4}
p-values	1.88×10^{-4}		4.19×10^{-3}		2.68×10^{-6}		8.56×10^{-8}	
Time (s)	0.47	0.61	0.75	0.99	0.79	0.86	0.93	1.17
NoI.	161	182	270	294	288	313	363	396



(a) (b)

(c) (d)

Fig. 4: ELBO for each iteration of each simulated dataset. The initial NoC is 15. Vertical lines indicate cancellation of components. (a) D1. (b) D2. (c) D3. (d) D4.



(a) (b)

(c) (d)

Fig. 5: Box plots for comparisons of the ELBOs' distributions for the InBLMM$_{\mathrm{SLB}}$ and the InBLMM$_{\mathrm{MLB}}$ with different simulated datasets. (a) D1. (b) D2. (c) D3. (d) D4.

histogram of oriented gradient (HOG) [66] descriptor has been one of the most popular ones for object detection, due to its promising results in many real-world applications [66]–[69]. Here, a rectangular HOG (R-HOG) descriptor [70] is applied, which is an improved version of the HOG. By considering 3 windows and 9 histogram bins for the R-HOG, we can represent each image with an 81-dimensional vector of features.

Experiments were conducted on four public available Caltech datasets[2] to evaluate our model for detecting some complex objects. These datasets are Caltech airplanes (1074 images), Caltech faces (450 images), Caltech car-sides (1155 images) and Caltech motorbikes (826 images) datasets. For non-object images, the Caltech background dataset (451 images) was also applied. Sample images from these datasets are displayed in Fig. 7. To avoid the randomness effects, each of the object datasets and the non-object dataset were randomly split into two halves, one for training and the other for testing. During the evaluation, we trained one detector for each object class. These procedure has been conducted 20 times.

With the above setting, the detection process in our case can be summarized as follows. First, the R-HOG descriptors were extracted from each image that is then represented as a positive feature vector. Second, the obtained feature vectors were normalized by dividing them by their l_1 norm, each image was thus finally represented as a proportional vector. Third, these proportional vectors were applied to train the classifiers for each object class, and each class was thus represented as an InBLMM. Finally, the test images were assigned to the group which had the highest posterior probability according to the Bayes's rule. For comparison, we have also applied four referred approaches for the object detection, including InBLMM$_{\mathrm{MLB}}$, InDMM$_{\mathrm{SLB}}$ [71], InDMM$_{\mathrm{MLB}}$ [53], and infinite Gaussian mixture learned in variational way

[2] http://www.vision.caltech.edu/html-files/archive.html.

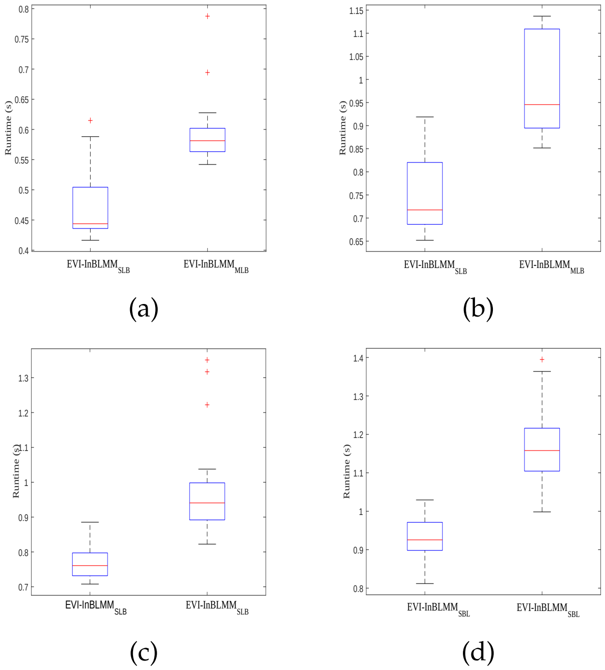


Fig. 6: Box plots for comparisons of the convergence runtime's distributions for the InBLMM$_{\text{SLB}}$ and the InBLMM$_{\text{MLB}}$ with different simulated datasets. (a) D1. (b) D2. (c) D3. (d) D4.

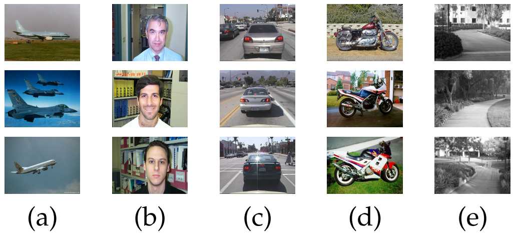


Fig. 7: Example images from the Caltech datasets. (a) Airplane. (b) Face. (c) Car-side. (d) Motorbike. (e) Background.

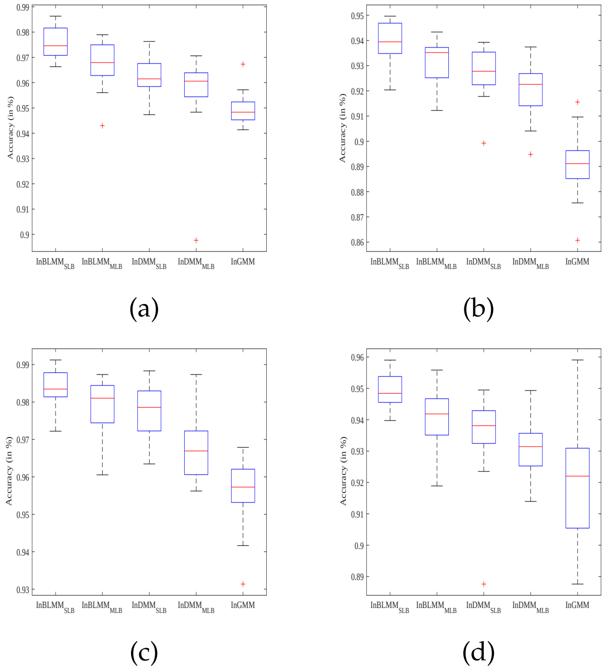


Fig. 8: The boxplots of the detection accuracies in the Caltech datasets. The central red mark is the median, the blue edges of the box are the 25th and 75th percentiles, respectively. (a) Airplane. (b) Face. (c) Car-side. (d) Motorbike.

(InGMM) [52]. It is noteworthy that the main motivation is to evaluate the methods in the object detection through considering comparable mixture-based methods. Hence, comparing our results to other non-mixture model-based methods is out of the scope of this paper. The average detection accuracies with the standard deviations for the aforementioned datasets are reported in Table 4. It is clearly presented that the SLB approximation (*i.e.*, InBLMM$_{\text{SLB}}$ and InDMM$_{\text{SLB}}$) performs better than the MLB approximations (*i.e.*, InBLMM$_{\text{MLB}}$ and InDMM$_{\text{MLB}}$), which is consistent with the previous study [27]. Fig. 8 shows the distributions of the detection accuracies.

4.2.2 *Text Categorization*

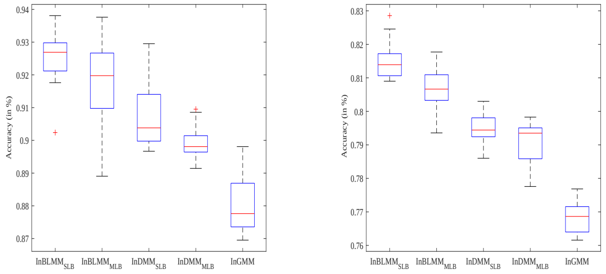
With the rapid development of information technology and popularization of the Internet, the number of the text documents increases exponentially. If the text documents are manually organized and managed only, it not only consumes much manpower and time, but also is hard to be achieved. As automatic text categorization (TC) has the realistic significance for efficient management and effective utilization of text information, it has become an active research topic in a variety of domains such as information retrieval, data mining, and statistical learning. During the past few years, a great number of statistical and machine learning methods have been proposed to address this problem [72].

In this section, the proposed InBLMM$_{\text{SLB}}$ is applied as a classifier for the TC task and the experimental results are reported on two publicly available datasets namely `WebKB` and `20Newsgroups`. The `WebKB` is composed of 4199 web pages from four categorizations namely course, faculty, project, and student. The `20Newsgroups` contains $13,998$ newsgroup documents, which are categorized into 20 different newsgroups. Each of these datasets was 20 times randomly split into two separate halves, one for training and the other one for testing. Following the work in [73], the Porters stemming [74] was used to reduce the words to their base forms. In the pre-processing stage, words occurring less than 3 times or shorter than 2 in length were eliminated, such that each document was represented by a positive vector. Then, these feature vectors are normalized into proportional vectors before employing the InBLMM$_{\text{SLB}}$. After this stage, each category in the training set was represented by an InBLMM$_{\text{SLB}}$. Finally, in the testing stage each test vector was assigned to a given category

TABLE 4: Object detection accuracies on the Caletch datasets. The standard deviations are in the brackets. The p-values of the student's t-test with the null hypothesis that InBLMM$_{\text{SLB}}$ and the referred methods have equal means but unknown variances are listed.

Dataset	Method	InBLMM$_{\text{SLB}}$	InBLMM$_{\text{MLB}}$	InDMM$_{\text{SLB}}$	InDMM$_{\text{MLB}}$	InGMM
Airplane	Accuracy (in %)	**97.82**(0.62)	96.77(0.89)	96.29(0.75)	95.87(0.15)	94.21(0.61)
	p-values	N/A	6.78×10^{-11}	3.12×10^{-12}	1.49×10^{-8}	2.41×10^{-15}
Face	Accuracy (in %)	**93.89**(0.88)	93.08(0.91)	92.75(0.23)	92.06(1.09)	89.04(1.19)
	p-values	N/A	2.06×10^{-3}	5.79×10^{-16}	2.42×10^{-12}	3.13×10^{-11}
Car-side	Accuracy (in %)	**98.38**(0.67)	97.91(0.41)	97.64(0.83)	96.72(0.77)	95.61(0.85)
	p-values	N/A	8.11×10^{-9}	6.06×10^{-7}	4.37×10^{-14}	9.32×10^{-7}
Motorbike	Accuracy (in %)	**94.92**(0.92)	94.14(0.58)	93.55(1.31)	93.09(0.85)	91.86(1.17)
	p-values	N/A	4.16×10^{-21}	8.34×10^{-15}	1.73×10^{-15}	7.03×10^{-11}

according to the Bayes's rule. With the same procedure mentioned above, four other statistical model, *i.e.*, the InBLMM$_{\text{MLB}}$, the InDMM$_{\text{SLB}}$, the InDMM$_{\text{MLB}}$ and the InGMM, were also applied. Table 5 summarizes the the average classification accuracy rates. Fig. 9 shows the distributions of the categorization accuracies. According to these results, it can be observed that the InBLMM$_{\text{SLB}}$ provides the best categorization accuracies compared to three other methods, which further demonstrates the advantage of the SLB approximation over the MLB approximation.



(a) (b)

Fig. 9: The boxplots of the categorization accuracies in the two datasets. (a) `WebKB`. (b) `20Newsgroups`.

5 Conclusions

In this paper, a variational Bayesian inference method was proposed for the InBLMM which is the cornerstone of non-parametric Bayesian statistics. Due to the nature of Bayesian non-parametric approaches, the InBLMM overcame the limitations of model selection existing in the BLMM. Meanwhile, through the Bayesian estimation, the InBLMM mitigated the overfitting and underfitting problems. To make the estimation feasible, the truncated representation of the DP and the EVI framework were further applied such that an analytically tractable solution can be derived. Experimental results on both synthetic and real datasets demonstrate that the proposed method achieved superior performance than the referred methods. Furthermore, the way of extending the proposed method is to be investigated, *e.g.*, considering the inclusion of a feature selection component within the EVI framework.

Acknowledgment

Appendix A

The expectation of the logarithm of the joint distribution in (12) can be written as

$$
\begin{aligned}
&\langle \ln p(\mathbf{X}, \boldsymbol{\Theta}) \rangle \\
&= \sum_{n=1}^{N} \sum_{m=1}^{\infty} \langle z_{nm} \rangle \Bigg\{ \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1-\lambda_m) \rangle \\
&\quad + \sum_{d=1}^{D} (\langle \alpha_{md} \rangle - 1) \ln x_{nd} + (\langle u_m \rangle - \sum_{d=1}^{D} \langle \alpha_{md} \rangle) \\
&\quad \times \ln(\sum_{d=1}^{D} x_{nd}) + (\langle v_m \rangle - 1)(1 - \sum_{d=1}^{D} x_{nd}) + \mathcal{R}_m + \mathcal{F}_m \Bigg\} \\
&\quad + \sum_{m=1}^{M} [(\eta_m - 1)\langle \ln(1-\lambda_m) \rangle] + \sum_{m=1}^{M} [(g_m - 1)\langle \ln u_m \rangle \\
&\quad - h_m \langle u_m \rangle + (p_m - 1)\langle \ln v_m \rangle - q_m \langle v_m \rangle] \\
&\quad + \sum_{m=1}^{M} \sum_{d=1}^{D} [(s_{md} - 1)\langle \ln \alpha_{md} \rangle - t_{md} \langle \alpha_{md} \rangle] + \text{Con.},
\end{aligned}
\tag{36}
$$

where we define

$$
\mathcal{R}_m = \left\langle \ln \frac{\Gamma(u_m + v_m)}{\Gamma(u_m)\Gamma(v_m)} \right\rangle, \mathcal{F}_m = \left\langle \ln \frac{\Gamma(\prod_{d=1}^{D} \alpha_{md})}{\prod_{d=1}^{D} \Gamma(\alpha_{md})} \right\rangle. \tag{37}
$$

Unfortunately, with the mathematical expression in (36) and by using the standard VI, it is infeasible to develop an analytically tractable solution to the variational posterior distribution. This is due to the fact that both $\mathcal{R}_m$ and $\mathcal{F}_m$ cannot be computed in a closed form. The EVI framework introduced in the previous section can be adopted to deal with this problem in an elegant way.

By following the principles of the EVI, we have to find two auxiliary functions which satisfy constraints as $\tilde{\mathcal{R}}_m \le \mathcal{R}_m$ and $\tilde{\mathcal{F}}_m \le \mathcal{F}_m$. In fact, we can select $\tilde{\mathcal{R}}_m$ and $\tilde{\mathcal{F}}_m$ as

$$
\begin{aligned}
\tilde{\mathcal{R}}_m &= \ln \frac{\Gamma(\bar{u}_m + \bar{v}_m)}{\Gamma(\bar{u}_m)\Gamma(\bar{v}_m)} + [\Psi(\bar{u}_m + \bar{v}_m) - \Psi(\bar{u}_m)] \\
&\quad \times (\langle \ln u_m \rangle - \ln \bar{u}_m)\bar{u}_m + [\Psi(\bar{u}_m + \bar{v}_m) - \Psi(\bar{v}_m)] \\
&\quad \times (\langle \ln v_m \rangle - \ln \bar{v}_m)\bar{v}_m,
\end{aligned}
\tag{38}
$$

TABLE 5: Text categorization accuracies on the three datasets. The standard deviations are in the brackets. The p-values of the student's t-test with the nullhypothesis that InBLMM$_{\text{SLB}}$ and the referred methods have equal means but unknown variances are listed.

Dataset	Method	InBLMM$_{\text{SLB}}$	InBLMM$_{\text{MLB}}$	InDMM$_{\text{SLB}}$	InDMM$_{\text{MLB}}$	InGMM
`WebKB`	Accuracy (in %)	**92.58**(0.79)	91.64(0.13)	90.63(0.17)	89.93(0.25)	88.06(0.93)
	p-values	N/A	4.65×10^{-6}	7.18×10^{-9}	1.39×10^{-4}	5.74×10^{-11}
`20Newsgroups`	Accuracy (in %)	**81.51**(0.52)	80.67(0.66)	79.46(0.48)	79.11(0.53)	76.83(0.47)
	p-values	N/A	4.86×10^{-9}	3.72×10^{-4}	5.28×10^{-7}	6.64×10^{-6}

$$\tilde{\mathcal{F}}_m = \ln \frac{\Gamma(\sum_{d=1}^{D} \bar{\alpha}_{md})}{\prod_{d=1}^{D} \Gamma(\bar{\alpha}_{md})} + \sum_{d=1}^{D} \left[\Psi(\sum_{k=1}^{D} \bar{\alpha}_{mk}) - \Psi(\bar{\alpha}_{md}) \right] \times \left[\langle \ln \alpha_{md} \rangle - \ln \bar{\alpha}_{md} \right] \bar{\alpha}_{md}. \tag{39}$$

For more additional derivation details for $\tilde{\mathcal{R}}_m$ and $\tilde{\mathcal{F}}_m$, please refer to [27]. In (38) and (39), $\Psi(\cdot)$ is the digamma function that is defined as $\Psi(a) = \partial \ln \Gamma(a) / \partial a$.

By substituting (38) and (39) back into (36), we obtain a lower bound to $\langle \ln p(\mathbf{X}, \mathbf{\Theta}) \rangle$ as

$$\begin{aligned} &\langle \ln \tilde{p}(\mathbf{X}, \mathbf{\Theta}) \rangle \\ =& \sum_{n=1}^{N} \sum_{m=1}^{\infty} \langle \mathrm{z}_{nm} \rangle \Bigg\{ \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1 - \lambda_m) \rangle \\ &+ \sum_{d=1}^{D} (\langle \alpha_{md} \rangle - 1) \ln \mathrm{x}_{nd} + (\langle u_m \rangle - \sum_{d=1}^{D} \langle \alpha_{md} \rangle) \\ &\times \ln(\sum_{d=1}^{D} \mathrm{x}_{nd}) + (\langle v_m \rangle - 1)(1 - \sum_{d=1}^{D} \mathrm{x}_{nd}) + \tilde{\mathcal{R}}_m + \tilde{\mathcal{F}}_m \Bigg\} \\ &+ \sum_{m=1}^{M} [(\eta_m - 1) \langle \ln(1 - \lambda_m) \rangle] + \sum_{m=1}^{M} [(g_m - 1) \langle \ln u_m \rangle \\ &- h_m \langle u_m \rangle + (p_m - 1) \langle \ln v_m \rangle - q_m \langle v_m \rangle] \\ &+ \sum_{m=1}^{M} \sum_{d=1}^{D} [(s_{md} - 1) \langle \ln \alpha_{md} \rangle - t_{md} \langle \alpha_{md} \rangle] + \text{Con.}. \end{aligned} \tag{40}$$

Applying (20) and with the joint distribution $\langle \ln \tilde{p}(\mathbf{X}, \mathbf{\Theta}) \rangle$ in (40), it is straightforward to develop analytically tractable solutions to the optimal posterior distributions. It is worth to note that in the following VI process, the stick-breaking representation for the InBMM is truncated at a level of M. The details with respect to the derivation of the updating equations for the hyperparamters are given as follows.

A. Proof for (26)*: Update* r_{nm}

By considering z_{nm} as the variable and absorbing any term that is independent of z_{nm} into the additional normalization constant, we have

$$\begin{aligned} \ln q(\mathrm{z}_{nm}) =& \mathrm{z}_{nm} \Bigg\{ \tilde{\mathcal{R}}_m + \tilde{\mathcal{F}}_m + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1 - \lambda_m) \rangle \\ &+ \sum_{d=1}^{D} (\langle \alpha_{md} \rangle - 1) \ln \mathrm{x}_{nd} + (\langle u_m \rangle - \sum_{d=1}^{D} \langle \alpha_{md} \rangle) \\ &\times \ln(\sum_{d=1}^{D} \mathrm{x}_{nd}) + (\langle v_m \rangle - 1) \ln(1 - \sum_{d=1}^{D} \mathrm{x}_{nd}) \Bigg\} + \text{Con.}. \end{aligned} \tag{41}$$

By taking a closer look at (41), it can be visualized that (41) has the logarithmic form of (7) except for the normalization constant. Therefore, $\ln q(\mathbf{Z})$ can be written in the form as

$$\ln q(\mathbf{Z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathrm{z}_{nm} \ln \rho_{nm} + \text{Con.}, \tag{42}$$

where ρ_{nm} has the form of (27). Taking logarithm on both sides in (41), we have

$$q(\mathbf{Z}) \propto \prod_{n=1}^{N} \prod_{m=1}^{M} \rho_{nm}^{\mathrm{z}_{nm}}. \tag{43}$$

Notice that for each value of n, the quantities z_{nm} in (43) are binary and sum to 1. By normalizing $q(\mathbf{Z})$, we obtain

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{m=1}^{M} r_{nm}^{\mathrm{z}_{nm}}, \tag{44}$$

which is a categorical distribution with

$$r_{nm} = \frac{\rho_{nm}}{\sum_{n=1}^{N} \rho_{nm}}, \tag{45}$$

where r_{nm} are nonnegative and have a unit sum. The quantities r_{nm} pay the role of responsibilities. For $q(\mathbf{Z})$, we have $\langle \mathrm{z}_{nm} \rangle = r_{nm}$.

B. Proof for (28)*: Update* a_m^* *and* b_m^*

Likewise, any terms that do not depend on λ_m will be absorbed into the additive constant, such that we have

$$\begin{aligned} \ln q(\lambda_m) =& \ln \lambda_m \sum_{n=1}^{N} \langle \mathrm{z}_{nm} \rangle + \ln(1 - \lambda_m) \\ &\times \left[\sum_{n=1}^{N} \sum_{j=m+1}^{M} \langle \mathrm{z}_{nj} \rangle + \eta_m - 1 \right] + \text{Con.}. \end{aligned} \tag{46}$$

We recognize this as the log of a Beta distribution, and so identifying the coefficients of $\ln\lambda_m$ and $\ln(1-\lambda_m)$. We obtain

$$q(\boldsymbol{\lambda}) = \prod_{m=1}^{M} \mathrm{Beta}(\lambda_m|a_m^*, b_m^*), \tag{47}$$

with the hyperparameters a_m^* and b_m^* given by (28).

C. Proof for (29) *and* (30)*: Update* s_{md}^* *and* t_{md}^*

Again, using the general result (20), and keeping only those terms that have a functional dependence on α_{md}, we have

$$\begin{aligned}\ln q(\alpha_{md}) = &\left\{\sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\left[\Psi(\sum_{l=1}^{D}\bar{\alpha}_{ml}) - \Psi(\alpha_{md})\right]\bar{\alpha}_{md}\right.\\ &\left.+s_{md}-1\right\}\ln\alpha_{md} - \left[t_{md} - \sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\ln x_{nd})\right]\alpha_{md}\\ &+\mathrm{Con.}.\end{aligned} \tag{48}$$

By taking exponential of both sides of (48), $q(\boldsymbol{\Lambda})$ is recognized to be a Gamma density

$$q(\boldsymbol{\Lambda}) = \prod_{m=1}^{M}\prod_{d=1}^{D}\mathcal{G}(\alpha_{md}|s_{md}^*, t_{md}^*), \tag{49}$$

where the optimal solutions to the hyperparameters s_{md}^* and t_{md}^* are given by (29) and (30), respectively.

D. Proof for (33) *and* (34)*: Update* g_m^* *and* h_m^*

The variational posterior $\ln q(u_m)$ is found by keeping terms related to u_m as

$$\begin{aligned}\ln q(u_m) = \mathrm{Con.} + &\left\{\sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\left[\Psi(\bar{u}_m+\bar{v}_m) - \Psi(\bar{u}_m)\right]\bar{u}_m\right.\\ &\left.+g_m-1\right\}\ln u_m - \left[h_m - \sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\ln(\sum_{d=1}^{D}\mathrm{x}_{nd})\right]u_m.\end{aligned} \tag{50}$$

It can be shown that (50) has the logarithmic form of a Gamma distribution as its conjugate prior distribution (10). By taking the exponential of its both sides, we then have

$$q(\mathbf{U}) = \prod_{m=1}^{M}\prod_{d=1}^{D}\mathcal{G}(u_m|g_m^*, h_m^*), \tag{51}$$

where the hyperparameters g_m^* and h_m^* are given by (33) and (34), respectively.

E. Proof for (35) *and* (36)*: Update* p_m^* *and* q_m^*

Similar to (50), the logarithm of the variational factor $q(v_m)$ can be calculated as

$$\begin{aligned}\ln q(v_m) = \mathrm{Con.} + &\left\{\sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\left[\Psi(\bar{u}_m+\bar{v}_m) - \Psi(\bar{v}_m)\right]\bar{v}_m\right.\\ &\left.+p_m-1\right\}\ln v_m - \left[q_m - \sum_{n=1}^{N}\langle \mathrm{z}_{nm}\rangle\ln(1-\sum_{d=1}^{D}\mathrm{x}_{nd})\right]v_m.\end{aligned} \tag{52}$$

It is obvious that (52) has the logarithmic form of a Gamma distribution as its conjugate prior distribution (11). By taking the exponential of its both sides, we obtain

$$q(\mathbf{V}) = \prod_{m=1}^{M}\prod_{d=1}^{D}\mathcal{G}(v_m|p_m^*, q_m^*), \tag{53}$$

where the hyperparameters p_m^* and q_m^* are given by (33) and (34), respectively. For further details on the derivation of variational learning, please refer to [3].

Appendix B

The expressions of the posterior expected values $\langle\cdot\rangle$ are given by

$$\langle \mathrm{z}_{nm}\rangle = r_{nm}, \langle\ln\lambda_m\rangle = \Psi(a_m) - \Psi(a_m+b_m), \tag{54}$$

$$\langle\ln(1-\lambda_m)\rangle = \Psi(b_m) - \Psi(a_m+b_m), \tag{55}$$

$$\langle\alpha_{md}\rangle = \bar{\alpha}_{md} = \frac{s_{md}^*}{t_{md}^*}, \langle\ln\alpha_{md}\rangle = \Psi(s_{md}^*) - \ln t_{md}^*, \tag{56}$$

$$\langle u_m\rangle = \bar{u}_m = \frac{g_m^*}{h_m^*}\ , \langle\ln u_{md}\rangle = \Psi(g_m^*) - \ln h_m^*, \tag{57}$$

$$\langle v_m\rangle = \bar{v}_m = \frac{p_m^*}{q_m^*}\quad \langle\ln v_{md}\rangle = \Psi(p_m^*) - \ln q_m^*. \tag{58}$$

Appendix C

The term $\langle\ln\tilde{p}(\mathbf{X},\boldsymbol{\Theta})\rangle$ in $\tilde{\mathcal{L}}(q)$ is given by (40), and the other terms $\langle\cdot\rangle$ are detailed as

$$\langle\ln q(\mathbf{Z})\rangle = r_{nm}\ln r_{nm}, \tag{59}$$

$$\begin{aligned}\langle\ln q(\boldsymbol{\lambda})\rangle = \sum_{m=1}^{M}&\left[\ln\Gamma(a_m^*+b_m^*) - \ln\Gamma(a_m^*) - \ln\Gamma(b_m^*)\right.\\ &\left.+\ (a_m^*-1)\langle\ln\lambda_m\rangle + (b_m^*-1)\langle\ln(1-\lambda_m)\rangle\right],\end{aligned} \tag{60}$$

$$\begin{aligned}\langle\ln q(\boldsymbol{\Lambda})\rangle = \sum_{m=1}^{M}\sum_{d=1}^{D}&\left[s_{md}^*\ln t_{md}^* - \ln\Gamma(s_{md}^*)\right.\\ &\left.+(s_{md}^*-1)\langle\ln\alpha_{md}\rangle - t_{md}^*\bar{\alpha}_{md}\right],\end{aligned} \tag{61}$$

$$\begin{aligned}\langle\ln q(\mathbf{U})\rangle = \sum_{m=1}^{M}&\left[g_m^*\ln h_m^* - \ln\Gamma(g_m^*) + (g_m^*-1)\right.\\ &\left.\times\langle\ln u_m\rangle - h_m^*\bar{u}_m\right],\end{aligned} \tag{62}$$

$$\begin{aligned}\langle\ln q(\mathbf{V})\rangle = \sum_{m=1}^{M}&\left[p_m^*\ln q_m^* - \ln\Gamma(p_m^*) + (p_m^*-1)\right.\\ &\left.\times\langle\ln v_m\rangle - q_m^*\bar{v}_m\right].\end{aligned} \tag{63}$$

References

[1] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[2] G. J. Mclachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[4] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1533–1543, 2004.

[5] D. A. Reynolds and R. C. Rose, "Robust text-ndependent speaker identification using gaussian mixture mpeaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[6] Y. Zhao and Y. Su, "Vehicles detection in complex urban scenes using gaussian mixture model with fmcw radar," *IEEE Sensors Journal*, vol. 17, no. 18, pp. 5948–5953, 2017.

[7] A. Juan and E. Vidal, "On the use of bernoulli mixture models for text classification," *Pattern Recognition*, vol. 35, no. 12, pp. 2705–2710, 2002.

[8] T. M. Nguyen and Q. M. J. Wu, "A nonsymmetric mixture model for unsupervised image segmentation," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 751–765, 2013.

[9] H. Zhang, Q. M. J. Nguyen, and X. Sun, "Synthetic aperture radar image segmentation by modified student's t-mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4391–4403, 2014.

[10] C. Constantinopoulos, M. K. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1013–1018, 2006.

[11] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, "Efficient volume exploration using the gaussian mixture model," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 11, pp. 1560–1573, 2011.

[12] K. Todros and J. Tabrikian, "Blind separation of independent sources using gaussian mixture model," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3645–3658, 2007.

[13] S. Zhang, L. Jiao, F. Liu, and S. Wang, "Global low-rank image restoration with gaussian mixture model," *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1827–1838, 2018.

[14] D. Peel and G. J. Mclachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 335–344, 2000.

[15] M. Svensen and C. M. Bishop, "Robust bayesian mixture modelling," *Neurocomputing*, vol. 6, no. 4, pp. 235–252, 2005.

[16] A. Matza and Y. Bistritz, "Skew gaussian mixture models for speaker recognition," *IET Signal Processing*, vol. 8, no. 8, pp. 860–867, 2014.

[17] S. X. Lee, K. L. Leemaqz, and G. J. McLachlan, "A block em algorithm for multivariate skew normal and skew t -mixture models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5581–5591, 2018.

[18] N. Bouguila, "Hybrid generative/discriminative approaches for proportional data modeling and classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2184–2202, 2012.

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

[20] Z. Ma, Y. Lai, K. W. Bastiaan, Y. Z. Song, W. Liang, and J. Guo, "Variational bayesian learning for dirichlet process mixture of inverted dirichlet distributions in non-gaussian image feature modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 449–463, 2019.

[21] 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, no. 3, pp. 269–284, 2019.

[22] J. Taghia and A. Leijon, "Variational inference for watson mixture model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1886–1900, 2016.

[23] W. Fan, N. Bouguila, J. Du, and X. Liu, "Axially symmetric data clustering through dirichlet process mixture models of watson distributions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1683–1694, 2019.

[24] J. Taghia, Y. Ma, and A. Leijon, "Bayesian estimation of the von-mises fisher mixture model with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, 2016.

[25] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2160–2173, 2011.

[26] W. Fan, N. Bouguila, and D. Ziou, "Variational learning for finite dirichlet mixture models and applications," *IEEE Transactions on Neural Network and Learning System*, vol. 23, no. 5, pp. 762–774, 2012.

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

[28] W. Fan and N. Bouguila, "Learning finite beta-liouville mixture models via variational bayes for proportional data clustering," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 1323–1329.

[29] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Annals of Statistics*, vol. 39, no. 1, pp. 1–38, 1974.

[31] D. C. Stanford and A. E. Raftery, "Approximate bayes factors for image segmentation: the pseudolikelihood information criterion (plic)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1517–1520, 2002.

[32] S. Favaro and Y. W. Teh, "Mcmc for normalized random measure mixture models," *Statistics*, vol. 28, no. 3, pp. 335–359, 2013.

[33] C. P. Robert, *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[34] N. Hjort, C. Holmes, P. Muller, and S. G. Waller, *Bayesian Non Parametrics: Principles and practice*. C. U. Press, Ed., 2010.

[35] C. E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[36] T. Chen, J. Morris, and E. Martin, "Probability density estimation via an infinite gaussian mixture model: Application to statistical process monitoring," *Journal of the Royal Statistical Society: Series C*, vol. 55, no. 5, pp. 699–715, 2006.

[37] M. Heck, S. Sakti, and S. Nakamura, "Dirichlet process mixture of mixtures model for unsupervised subword modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2027–2042, 2018.

[38] C. E. Rasmussen, "The infinite gaussian mixture model," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1999, pp. 554–560.

[39] W. Fan and N. Bouguila, "Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference," *IEEE Transactions on Neural Network and Learning Systems*, vol. 24, no. 11, pp. 1850–1862, 2013.

[40] Y. Lai, Y. Ping, K. Xiao, B. Hao, and X. Zhang, "Variational bayesian inference for a dirichlet process mixture of beta distributions and application," *Neurocomputing*, vol. 278, no. 2, pp. 23–33, 2017.

[41] C. Andrieu, N. D. Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[42] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[43] Y. Lai, P. Yuan, W. He, B. Wang, J. Wang, and 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.

[44] Z. Ma, J. Xie, Y. Lai, D. Meng, W. B. Kleijn, and 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 Systems*, vol. 31, no. 7, pp. 2240–2254, 2020.

[45] N. Bouguila, “Bayesian hybrid generative discriminative learning based on finite liouville mixture models,” *Pattern Recognition*, vol. 44, no. 6.

[46] ——, “On the smoothing of multinomial estimates using liouville mixture models and applications,” *Pattern Analysisi and Application*, vol. 16, no. 3.

[47] K. T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions*. London, U.K.: Chapman and Hal, 1990.

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

[49] T. S. Ferguson, “A bayesian analysis of some nonparametric problems,” *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[50] J. Sethuraman, “A constructive definition of dirichlet priors,” *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[51] H. Ishwaran and L. F. James, “Gibbs sampling methods for stick-breaking priors,” *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[52] D. M. Blei and M. I. Jordan, “Variational inference for dirichlet process mixtures,” *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2005.

[53] W. Fan and N. Bouguila, “Variational learning for dirichlet process mixtures of dirichlet distributions and applications,” *Multimedia Tools and Applications*, vol. 70, no. 3, pp. 1685–1702, 2014.

[54] ——, “Variational learning for dirichlet process mixtures of dirichlet distributions and applications,” *Multimedia Tools and Applications*, vol. 70, no. 8, pp. 1685–1702, 2014.

[55] A. Corduneanu and C. Bishop, “Variational bayesian model selection for mixture distributions,” in *Proceedings of Eighth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2001, pp. 27–34.

[56] A. Andreopoulos and J. K. Tsotsos, “50 years of object recognition: Directions forward,” *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 827–891, 2013.

[57] K. Oh, M. Lee, Y. Lee, and S. Kim, “Salient object detection using recursive regional feature clustering,” *Information Science*, vol. 387, no. C, pp. 1–18, 2017.

[58] B. H. Chen and S. C. Huang, “An advanced moving object detection algorithm for automatic traffic monitoring in real-world limited bandwidth networks,” *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 837–847, 2014.

[59] C. R. Del-Blanco, F. Jaureguizar, and N. Garcia, “An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications,” *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 857–862, 2012.

[60] T. H. Tsai, C. Y. Lin, and S. Y. Li, “Algorithm and architecture design of humancmachine interaction in foreground object detection with dynamic scene,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 15–29, 2013.

[61] T. Lei and J. Udupa, “A sensor array processing approach to object region detection.” *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1319–1325, 2001.

[62] Y. S. Chia, S. Rahardja, D. Rajan, and M. K. H. Leung, “Structural descriptors for category level object detection,” *IEEE Transactions on Multimedia*, vol. 11, no. 8, pp. 1407–1421, 2009.

[63] X. Bai, H. Zhang, and J. Zhou, “Vhr object detection based on structural feature extraction and query expansion,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6508–6520, 2014.

[64] G. C. Chen and C. F. Juang, “Object detection using color entropies and a fuzzy classifier,” *IEEE Computational Intelligence Magazine*, vol. 8, no. 1, pp. 33–45, 2013.

[65] Y. Li, S. Wang, Q. Tian, and X. Ding, “Learning cascaded shared-boost classifiers for part-based object detection,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1858–1871, 2014.

[66] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[67] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, “Airborne vehicle detection in dense urban areas using hog features and disparity maps,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2327–2337, 2013.

[68] S. Bourouis, M. A. Mashrgy, and 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.

[69] C. G. Blair and N. M. Robertson, “Video anomaly detection in real time on a power-aware heterogeneous platform,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2109–2122, 2016.

[70] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, “Trainable classifier-fusion schemes: An application to pedestrian detection,” in *IEEE Int’l Conf. Intelligent Transportation Systems (ITS)*, 2009, pp. 1–6.

[71] Y. Lai, W. He, Y. Ping, J. Qu, and X. Zhang, “Variational bayesian inference for infinite dirichlet mixture towards accurate data categorization,” *Wireless Personal Communications*, vol. 102, no. 3, p. 2307C2329, 2018.

[72] D. Cai and X. He, “Manifold adaptive experimental design for text categorization,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 4, pp. 707–719, 2012.

[73] Y. Ping, Y. Tian, C. Guo, B. Wang, and Y. Yang, “Frsvc: Towards making support vector clustering consume less,” *Pattern Recognition*, vol. 69, no. 9, pp. 286–298, 2017.

[74] M. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.

Yuping Lai is an associate professor of the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China. He received his Ph.D. degree in Information Security from Beijing University of Posts and Telecommunications, Beijing, China, in 2014. His research interests include information security, computer vision, pattern recognition, machine learning, and data mining.

Lijuan Luo has been an assistant professor at Shanghai International Studies University, China, since 2015. He received her Ph.D. degree in Management Science and Engineering from Beijing University of Posts and Telecommunications, China, in 2015. Her research interests include information management, machine learning, big data and data decision.

Yuan Ping received the PhD degree from Beijing University of Posts and Telecommunications, China, in 2012. He was a research assistant in University of Alberta from September 2014 to August 2015. He is currently an Associate Professor, and the Director of Network and Information Security Lab of Xuchang University. His research interests are in machine learning, data mining, pattern recognition, information security, and operating system.

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.

Hongying Meng received the Ph.D. degree in communication and electronic systems from Xi’an Jiaotong University. He was a Lecturer with the Electronic Engineering Department, Tsinghua University, China. He is currently a Reader with the Electronic and Electrical Engineering Department, Brunel University London, U.K. He has wide research interests, including digital signal processing, machine learning, human-computer interaction, computer vision, image processing, and embedded systems.