Variational Bayesian Learning for Dirichlet Process Mixture of Inverted Dirichlet Distributions

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID	TNNLS-2018-P-8972
Manuscript Type:	Paper
Date Submitted by the Author:	21-Jan-2018
Complete List of Authors:	Ma, Zhanyu; Beijing University of Posts and Telecommunications, Pattern Recognition and Machine Intelligenc Lab. Lai, Yuping; North China University of Technology, Department of Information Security Kleijn, W. Bastiaan; Victoria University of Wellington, School of Engineering and Computer Science Song, YiZhe; Queen Mary University of London, Wang, Liang; CASIA, NLPR Guo, Jun; Beijing University of Posts and Telecommunications, Pattern Recognition and Machine Intelligenc Lab.
Keywords:	Dirichlet process mixture, inverted Dirichlet distribution, Bayesian estimation, variational learning, computer vision

Variational Bayesian Learning for Dirichlet Process Mixture of Inverted Dirichlet Distributions

Zhanyu Ma, *Senior Member, IEEE*, Yuping Lai, *Member, IEEE*, W. Bastiaan Kleijn, *Fellow, IEEE*, Yi-Zhe Song, *Member, IEEE*, Liang Wang, *Senior Member, IEEE*, and Jun Guo

***Abstract*—In this work, we develop a novel variational Bayesian learning method for the Dirichlet process (DP) mixture of the inverted Dirichlet distributions, which has been shown to be very flexible for modeling vectors with positive elements. The recently proposed extended variational inference (EVI) framework is adopted to derive an analytically tractable solution. The convergency of the proposed algorithm is theoretically guaranteed by introducing single lower bound approximation to the original objective function in the EVI framework. In principle, the proposed model can be viewed as an infinite inverted Dirichelt mixture model (InIDMM) that allows the automatic determination of the number of mixture components from data. Therefore, the problem of pre-determining the optimal number of mixing components has been overcome. Moreover, the problems of over-fitting and under-fitting are avoided by the Bayesian estimation approach. Comparing with several recently proposed DP-related methods and conventional applied methods, the good performance and effectiveness of the proposed method have been demonstrated with both synthesized data and real data evaluations.**

***Index Terms*—Dirichlet process mixture, inverted Dirichlet distribution, Bayesian estimation, variational learning, computer vision**

I. Introduction

Finite mixture modeling [1], [2] is a flexible and powerful probabilistic modeling tool for data that are assumed to be generated from heterogeneous populations. It has been widely applied to many areas, such as pattern recognition, machine learning, data mining, computer vision [3]–[7]. Among all finite mixture models, the finite Gaussian mixture model (GMM) has been the most popular method for modeling continuous data. Much of its popularity is due to the fact that any continuous distribution can be arbitrarily well approximated by a GMM with unlimited number of mixture components. Moreover, the parameters in a GMM can be estimated efficiently via maximum likelihood (ML) estimation with the expectation maximum (EM) algorithm [8]. By assigning prior distributions to the parameters in a GMM, Bayesian estimation of GMM can be carried out with conjugate prior-posterior pair matching [9], [10]. Both the ML and the Bayesian estimation algorithms can be represented in analytically tractable form [9].

Recent studies have shown that non-Gaussian statistical models, *e.g.*, the beta mixture model (BMM) [6], the Dirichlet mixture model (DMM) [7], the Gamma mixture model (GaMM) [11], the von Mises-Fisher mixture model (vMM) [12], can model the non-Gaussian distributed data more efficiently, compared to the conventional GMM. For example, BMM has been widely applied in modeling grey image pixel values [6] and DNA methylation data [13]. In order to efficiently model proportional data [7], [14], DMM can be utilized to describe the underlying distribution. In generalized-K (K_G) fading channels, GaMM has been used to analyze the capacity and error probability [11]. The vMM has been widely used in modeling directional data, such as yeast gene expression [12] and topic detection [15]. The finite inverted Dirichlet mixture model (IDMM), among others, has been demonstrated to be an efficient tool for modeling data vector with positive elements [16], [17]. Moreover, the inverted Dirichlet distribution also has connections with nonnegative matrix factorization (NMF). In sparse NMF [18], the l_1-norm constraint is usually applied to favor the sparseness. As the definition of the inverted Dirichlet distribution is similar to the nonnegative properties of the columns in the original matrix and the basis matrix, selecting proper prior distribution to describe the underlying distribution of the aforementioned columns can favor the sparse NMF.

An essential problem in finite mixture modeling is how to automatically decide the appropriate number of mixture components based on the data. The component number has a strong effect on the modeling accuracy [19]. If the number of mixture components is not properly chosen, the mixture model may over-fit or under-fit the observed data. To deal with this problem, many methods have been proposed. These can be categorized into two groups: deterministic approaches [20], [21] and Bayesian methods [22], [23]. Deterministic approaches are generally implemented by ML estimation under an EM-based and require the integration of entropy measures or some information theoretic criteria, such as the minimum message length (MML) [21], the Bayesian information criterion (BIC) [24], and the Akaike information criterion (AIC) [25], to determine the number of components in the mixture model. It is worth noting that, in general, the EM algorithm converges to a local maximum or a saddle point and its solution is highly dependent on its initialization. On the other hand, the Bayesian methods, which are not sensitive to initialization

Z. Ma and J. Guo are with the Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, Beijing, China.

Y. Lai is with the Department of Information Security, North China University of Technology, Beijing, China.

W. B. Kleijn is with the Communications and Signal Processing Group, Victoria University of Wellington, New Zealand.

Y.-Z. Song is with the SketchX Lab, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK.

L. Wang is with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

The corresponding authors are Z. Ma (mazhanyu@bupt.edu.cn) and Y. Lai (laiyp@ncut.edu.cn).

by introducing proper prior distributions to the parameters in the model, have been widely used to find a suitable number of components in a finite mixture model. In this case, the parameters of a finite mixture model (including the parameters in a component and the weighting coefficients) are treated as random variables under the Bayesian framework. The posterior distributions of the parameters, rather than simple point estimates, are computed [2]. The model truncation in Bayesian estimation of finite mixture model is carried out by setting the corresponding weights of the unimportant mixture components to zero (or a small value close to zero) [2]. However, the number of mixture components should be properly initialized, as it can only decrease during the training process.

The increasing interest in mixture modeling has led to the development of the model selection method[1]. Recent work has shown that the non-parametric Bayesian approach [26]–[30] can provide an elegant solution for automatically determining the complexity of model. The basic idea behind this approach is that it provides methods to adaptively select the optimal number of mixing components, while also allows the number of mixture components to remain unbounded. In other words, this approach allows the number of components to increase as new data arrives, which is the key difference from finite mixture modeling. The most widely used Bayesian nonparametric [31] model selection method is based on the Dirichlet process (DP) mixture model [32], [33]. The DP mixture model extends distributions over measures, which has the appealing property that it does not need to set a prior on the number of components. In essence, the DP mixture model can also be viewed as an infinite mixture model with its complexity increasing as the size of dataset grows. Recently, the DP mixture model has been applied in many important applications. For instance, the DP mixture model has been adopted to a mixture of different types of non-Gaussian distributions, such as the DP mixture of beta-Liouville distributions [34], the DP mixture of student's-t distributions [35], the DP mixture of generalized Dirichlet distributions [36], the DP mixture of student's-t factors [37], and the DP mixture of hidden Markov random field models [38].

Generally speaking, most parameter estimation algorithms for both the deterministic and the Bayesian methods are time consuming, because they have to numerically evaluate a given model selection criterion [21]. This is especially true for the fully Bayesian Markov chain Monte Carlo (MCMC) [27], [39], which is one of the widely applied Bayesian approaches with numerical simulations. The MCMC approach has its own limitations, when high-dimensional data are involved in the training stage [40]. This is due to the fact that its sampling-based characteristics yield a heavy computational burden and it is difficult to monitor the convergence in the high-dimensional space. To overcome the aforementioned problems, variational inference (VI), which can provide an analytically tractable solution and good generalization performance, has been proposed as an efficient alternative to the MCMC approach [41]. With an analytically tractable solution, the numerical sampling during each iteration in the optimization stage can be avoided. Hence, the VI-based solutions can lead to more efficient estimation. They have been successfully applied in a variety of applications including the estimation of mixture models [5]–[7], [34], [42].

Motivated by the ability of the Bayesian non-parametric approaches to solve the model selection problem and the good performance recently obtained by the VI framework, we focus on the variational learning of the DP mixture of inverted Dirichlet distributions (*a.k.a.* the infinite inverted Dirichlet mixture model (InIDMM)). Since InIDMM is a typical non-Gaussian statistical model, it is not feasible to apply the standard VI framework to obtain an analytically tractable solution for the Bayesian estimation. As a variate of VI, stochastic variational infernece (SVI) [43], [44] has been proposed as an alternative solution to approximate the posterior distributions. The algorithm under SVI framework is scalable and suitable for massive data. However, when dealing with non-Gaussian distributions, the expectations in the update iterations (Fig. 4, [43]) cannot be calculated explicitly and some sampling methods are also required to approximate the expectations. In order to derive an analytically tractable solution for the variational learning of InIDMM, the recently proposed extended variational inference (EVI) [6], [7], which is particularly suitable for non-Gaussian statistical models, has been adopted to provide an appropriate *single lower bound (SLB) approximation* to the original object function. With the auxiliary function, an analytically tractable solution for Bayesian estimation of InIDMM is derived. The key contributions of our work are three-fold: 1) The finite inverted Dirichlet mixture model (IDMM) has been extended to the infinite inverted Dirichlet mixture model (InIDMM) under the stick-breaking process framework [32], [45]. Thus, the difficulty in automatically determining the number of mixture components can be overcome. 2) An analytically solution is derived with the EVI framework for InIDMM. Moreover, comparing with the recently proposed algorithm for InIDMM [46], which is based on *multiple lower bound (MLB) approximation*, our algorithm can not only theoretically guarantee convergence but also provide better approximations. 3) The proposed method has been applied in several important applications in computer vision, such as image categorization and object detection. The good performance has been illustrated with both synthesized and real data evaluations.

The remaining part of this paper is organized as follow: Section II provides a brief overview of the finite inverted Dirichlet mixture and the DP mixture. The infinite inverted Dirichlet mixture model is also proposed. In Section III, a Bayesian learning algorithm with EVI is derived. The proposed algorithm has an analytically tractable form. The experimental results with both synthesized and real data evaluations are reported in Section IV. Finally, we draw conclusions and future research directions in Section V.

II. The statistical model

In this section, we first present a brief overview of the finite inverted Dirichlet mixture model (IDMM). Then, the DP mix-

[1] Here, model selection means selecting the best of a set of models of different orders

ture model with stick-breaking representation is introduced. Finally, we extend the IDMM to InIDMM.

A. Finite inverted Dirichlet mixture model

Given a D-dimensional vector $\vec{x} = \{x_1, \cdots, x_D\}$ generated from an IDMM with M components, the probability density function (PDF) of $\vec{x}$ is denoted as [16]

$$\mathrm{IDMM}(\vec{x}|\vec{\pi}, \Lambda) = \sum_{m=1}^{M} \pi_m \mathrm{iDir}(\vec{x}|\vec{\alpha}_m), \tag{1}$$

where $\Lambda = \{\vec{\alpha}_m\}_{m=1}^{M}$ and $\vec{\pi} = \{\pi_m\}_{m=1}^{M}$ is the mixing coefficient vector subject to the constraints $0 \leq \pi_m \leq 1$ and $\sum_{m=1}^{M} \pi_m = 1$. Moreover, $\mathrm{iDir}(\vec{x}|\vec{\alpha})$ is an inverted Dirichlet distribution with its $(D+1)$-dimensional positive parameter vector $\vec{\alpha} = \{\alpha_1, \cdots, \alpha_{D+1}\}$ defined as

$$\mathrm{iDir}(\vec{x}|\vec{\alpha}) = \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1} \left(1 + \sum_{d=1}^{D} x_d\right)^{-\sum_{d=1}^{D+1} \alpha_d}, \tag{2}$$

where $x_d > 0$ for $d = 1, \cdots, D$ and $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$.

B. Dirichlet Process with Stick-Breaking

The Dirichlet process (DP) [32], [33] is a stochastic process used for Bayesian nonparametric data analysis, particularly in a DP mixture model (infinite mixture model). It is a distribution over distributions rather than parameters, *i.e.*, each draw from a DP is a probability distribution itself, rather than a parameter vector [47]. We adopt the DP to extend the IDMM to the infinite case, such that the difficulty of the automatic determination of the model complexity (*i.e.*, the number of mixture components) can be overcome. To this end, the DP is constructed by the following stick-breaking formulation [31], [48], [49], which is an intuitive and simple constructive definition of the DP.

Assume that H is a random distribution and φ is a positive real scalar. We consider two countably infinite collections of independently generated stochastic variables $\Omega_m \sim H$ and $\lambda_m \sim \mathrm{Beta}(\lambda_m; 1, \varphi)$[2] for $m = \{1, \cdots, \infty\}$, where $\mathrm{Beta}(x; a, b)$ is the beta distribution defined as $\mathrm{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$. A distribution G is said to be DP distributed with a concentration parameter φ and a base measure or base distribution H (denoted as $G \sim \mathrm{DP}(\varphi, H)$), if the following conditions are satisfied:

$$G = \sum_{m=1}^{\infty} \pi_m \delta_{\Omega_m}, \ \pi_m = \lambda_m \prod_{l=1}^{m-1} (1 - \lambda_l), \tag{3}$$

where $\{\pi_m\}$ is a set of stick-breaking weights with constraints $\sum_{m=1}^{\infty} \pi_m = 1$, δ_{Ω_m} is a delta function whose value is 1 at location Ω_m and 0 otherwise. The generation of the mixing coefficients $\{\pi_m\}$ can be considered as process of breaking a unit length stick into an infinite number of pieces. The length of each piece, λ_m, which is proportional to the rest of the "stick" before the current breaking, is considered as an independent random variable generated from $\mathrm{Beta}(\lambda_m; 1, \varphi)$. Because of its simplicity and natural generalization ability, the stick-breaking construction has been a widely applied scheme for the inference of DPs [34], [45], [50].

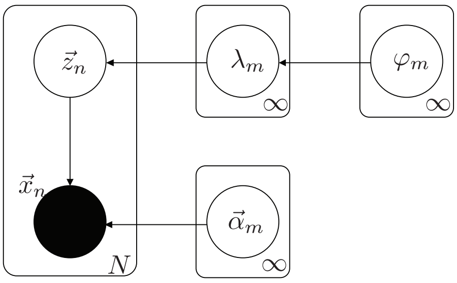


Fig. 1: Graphical representation of the variables relationships in the Bayesian inference of a InIDMM. All of the circles in the graphical figure represent variables. Arrows show the relationships between variables. The variables in the box are the *i.i.d.* observations.

C. Infinite Inverted Dirichlet Mixture Model

Now we consider the problem of modeling $\vec{x}$ by an Infinite Inverted Dirichlet Mixture Model (InIDMM), which is actually an extended IDMM with an infinite number of components. Therefore, (1) can be reformulated as

$$\mathrm{InIDMM}(\vec{x}|\vec{\pi}, \Lambda) = \sum_{m=1}^{\infty} \pi_m \mathrm{iDir}(\vec{x}|\vec{\alpha}_m), \tag{4}$$

where $\vec{\pi} = \{\pi_m\}_{m=1}^{\infty}$ and $\Lambda = \{\vec{\alpha}_m\}_{m=1}^{\infty}$. Then, the likelihood function of the InIDMM given the observed dataset $\mathcal{X} = \{\vec{x}_n\}_{n=1}^{N}$ is given by

$$\mathrm{InIDMM}(\mathcal{X}|\vec{\pi}, \Lambda) = \prod_{n=1}^{N} \left\{ \sum_{m=1}^{\infty} \pi_m \mathrm{iDir}(\vec{x}_n|\vec{\alpha}_m) \right\}. \tag{5}$$

In order to clearly illustrate the generation process of each observation $\vec{x}_n$ in the mixture model, we introduce a latent indication vector variable $\vec{z}_n = \{z_{n1}, z_{n2}, \cdots\}$. $\vec{z}$ has only one element equal to 1 and the other elements in $\vec{z}$ are 0. For example, $z_{nm} = 1$ indicates the sample $\vec{x}_n$ comes from the mixture component m. Therefore, the conditional distribution of $\mathcal{X}$ given the parameters Λ and the latent variables $\mathcal{Z} = \{z_{nm}\}$ is

$$\mathrm{InIDMM}(\mathcal{X}|\mathcal{Z}, \Lambda) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \mathrm{iDir}(\vec{x}_n|\vec{\alpha}_m)^{z_{nm}}. \tag{6}$$

Moreover, to exploit the advantages of the Bayesian framework, conjugate prior distributions are introduced for all the unknown parameters according to their distribution properties. In this work, we place the conjugate priors over the unknown stochastic variables $\mathcal{Z}$, Λ, and $\vec{\lambda} = (\lambda_1, \lambda_2, \cdots)$ such that a full Bayesian estimation model can be obtained.

In the aforementioned full Bayesian model, the prior distribution of $\mathcal{Z}$ given $\vec{\pi}$ is given by

$$p(\mathcal{Z}|\vec{\pi}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \pi_m^{z_{nm}}. \tag{7}$$

[2] To avoid confusion, we use $f(x; a)$ to denote the PDF of x parameterized by parameter a. $f(x|a)$ is used to denote the conditional PDF of x given a, where both x and a are random variables. Both $f(x; a)$ and $f(x|a)$ have exactly the same mathematical expressions.

As $\vec{\pi}$ is a function of $\vec{\lambda}$ according to the stick-breaking construction of the DP as shown in (3), we rewrite (7) as

$$p(\mathcal{Z}|\vec{\lambda}) = \prod_{n=1}^{N}\prod_{m=1}^{\infty}\left[\lambda_m \prod_{l=1}^{m-1}(1-\lambda_l)\right]^{z_{nm}}. \quad (8)$$

As previously mentioned in Section II-B, the prior distribution of $\vec{\lambda}$ is

$$p(\vec{\lambda}|\vec{\varphi}) = \prod_{m=1}^{\infty}\text{Beta}(\lambda_m; 1, \varphi_m) = \prod_{m=1}^{\infty}\varphi_m(1-\lambda_m)^{\varphi_m - 1}, \quad (9)$$

where $\vec{\varphi} = (\varphi_1, \varphi_2, \cdots)$. Based on (3), we can obtain the expected value of π_m. In order to do this, the expected value of λ_m will first be calculated as

$$\langle\lambda_m\rangle = 1/(1+\varphi_m). \quad (10)$$

Then, the expected value of π_m is denoted as

$$\langle\pi_m\rangle = \langle\lambda_m\rangle\prod_{l=1}^{m-1}(1-\langle\lambda_l\rangle). \quad (11)$$

It is worth to note that, when the value of φ_m is small, $\langle\lambda_m\rangle$ will become large. Therefore, the expected of the mixing coefficients π_m are controlled by the parameters φ_m, *i.e.*, small value of φ_m will yield small π_m such that the distribution of π_m will be sparse.

As φ_m is positive, we assume $\vec{\varphi}$ follows a product of gamma prior distributions as

$$p(\vec{\varphi}; \vec{s}, \vec{t}) = \prod_{m=1}^{\infty}\text{Gam}(\varphi_m; s_m, t_m) = \prod_{m=1}^{\infty}\frac{t_m^{s_m}}{\Gamma(s_m)}\varphi_m^{s_m - 1}e^{-t_m\varphi_m}, \quad (12)$$

where $\text{Gam}(\cdot)$ is the gamma distribution. $\vec{s} = (s_1, s_2, \cdots)$ and $\vec{t} = (t_1, t_2, \cdots)$ are the hyperparamters and subject to the constraints $s_m > 0$ and $t_m > 0$.

Next, we introduce an approximating conjugate prior distribution to parameter Λ in InIDMM. The inverted Dirichlet distribution belongs to the exponential family and its formal conjugate prior can be derived with the Bayesian rule [2] as

$$p(\vec{\alpha}|\vec{\mu}_0, v_0) = C(\vec{\mu}_0, v_0)\left[\frac{\Gamma(\sum_{d=1}^{D+1}\alpha_d)}{\prod_{d=1}^{D+1}\alpha_d}\right]^{\nu_0} e^{-\vec{\mu}_0(\vec{\alpha}^T - \vec{I}_{D+1})}, \quad (13)$$

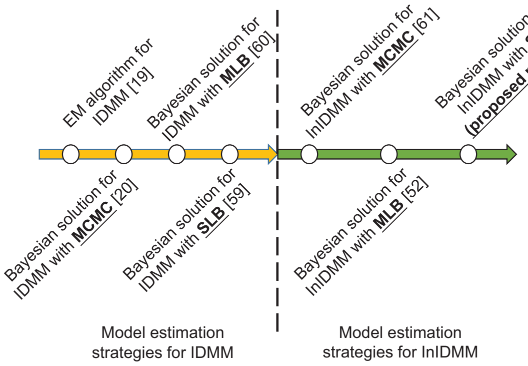
where $\vec{\mu}_0 = [\mu_{1_0}, \cdots \mu_{D+1_0}]$ and ν_0 are the hyperparameters in the prior distribution, $C(\vec{\mu}_0, v_0)$ is a normalization coefficient such that $\int p(\vec{\alpha}|\vec{\mu}_0, v_0)d\vec{\alpha} = 1$. $\vec{I}_d$ is a D-dimensional vector with all elements equal to one. Then, we can write the posterior distribution of $\vec{\alpha}$ as (with N *i.i.d.* observations $\mathcal{X}$)

$$\begin{aligned} f(\vec{\alpha}|\mathcal{X}) =& \frac{\text{iDir}(\mathcal{X}|\vec{\alpha})f(\vec{\alpha}|\vec{\mu}_0, \nu_0)}{\int \text{iDir}(\mathcal{X}|\vec{\alpha})f(\vec{\alpha}|\vec{\mu}_0, \nu_0)d\vec{\alpha}} \\ =& C(\vec{\mu}_N, \nu_N)\left[\frac{\Gamma(\sum_{d=1}^{D+1}\alpha_d)}{\prod_{d=1}^{D+1}\Gamma(\alpha_d)}\right]^{\nu_N} e^{-\vec{\mu}_N(\vec{\alpha}^T - \vec{I}_{D+1})} \end{aligned} \quad (14)$$

where the hyperparameters ν_N and $\vec{\mu}_N$ in the posterior distribution are

$$\nu_N = \nu_0 + N, \vec{\mu}_N = \vec{\mu}_0 - [\ln\mathcal{X}^+ - \vec{I}_{D+1}\ln(1 + \vec{I}_{D+1}^T\mathcal{X}^+)]\vec{I}_N. \quad (15)$$

In (15), $\mathcal{X}^+$ is a $(D+1)\times N$ matrix by connecting $\vec{I}_{D+1}^T$ to the bottom of $\mathcal{X}$. However, it is not applicable in our VI framework due to the analytically intractable normalization factor in (44). Because Λ is positive, we adopt gamma prior distributions to approximate conjugate prior for Λ as well. By assuming the parameters of inverted Dirichlet distribution are mutually independent, we have



EM algorithm for IDMM [19]
Bayesian solution for IDMM with MCMC [20]
Bayesian solution for IDMM with MLB [60]
Bayesian solution for IDMM with SLB [59]
Bayesian solution for InIDMM with MCMC [61]
Bayesian solution for InIDMM with MLB [52]
Bayesian solution for InIDMM with SLB (proposed method)
Model estimation strategies for IDMM
Model estimation strategies for InIDMM

Fig. 2: Development progress of the model estimation strategies for finite IDMM and infinite IDMM.

$$p(\Lambda) = \text{Gam}(\Lambda; U, V) = \prod_{m=1}^{\infty}\prod_{d=1}^{D+1}\frac{v_{md}^{u_{md}}}{\Gamma(u_{md})}\alpha_{md}^{u_{md}-1}e^{-v_{md}\alpha_{md}}, \quad (16)$$

where all the hyperparameters $U = \{u_{md}\}$ and $V = \{v_{md}\}$ are positive.

With the Bayesian rules and by combining (6) and (8)-(16) together, we can represent the joint density of the observation $\mathcal{X}$ with all the $i.i.d.$ latent variables $\Theta = (\mathcal{Z}, \Lambda, \vec{\lambda}, \vec{\varphi})$ as

$$\begin{aligned} p(\mathcal{X}, \Theta) =& p(\mathcal{X}|\mathcal{Z}, \Lambda)p(\mathcal{Z}|\vec{\lambda})p(\vec{\lambda}|\vec{\varphi})p(\vec{\varphi})p(\Lambda) \\ =& \prod_{n=1}^{N}\prod_{m=1}^{\infty}\Bigg[\lambda_m\prod_{j=1}^{m-1}(1-\lambda_j)\frac{\Gamma\left(\sum_{d=1}^{D+1}\alpha_{md}\right)}{\prod_{d=1}^{D+1}\Gamma(\alpha_{md})} \\ &\times\prod_{d=1}^{D}x_{nd}^{\alpha_{md}-1}\left(1+\sum_{d=1}^{D}x_{nd}\right)^{-\sum_{d=1}^{D+1}\alpha_{md}}\Bigg]^{z_{nm}} \\ &\times\prod_{m=1}^{\infty}\left[\varphi_m(1-\lambda_m)^{\varphi_m - 1}\frac{t_m^{s_m}}{\Gamma(s_m)}\varphi_m^{s_m-1}e^{-t_m\varphi_m}\right] \\ &\times\prod_{m=1}^{\infty}\prod_{d=1}^{D+1}\frac{v_{md}^{u_{md}}}{\Gamma(u_{md})}\alpha_{md}^{u_{md}-1}e^{-v_{md}\alpha_{md}}. \end{aligned} \quad (17)$$

The structure of the InIDMM can be represented in terms of a graphical model in Fig. 1. The development progress for the related models are shown in Fig. 2.

III. Variational Learning for InIDMM

In this section, we develop a variational Bayesian inference framework for learning the InIDMM. With the assistance of recently proposed EVI [6], [7], an analytically tractable algorithm, which prevents numerical sampling during each iteration and facilitates a training procedure, is obtained. The proposed solution is also able to overcome the problem of overfitting and automatically decide the number of mixture components.

A. Extended Variational Inference

The purpose of Bayesian analysis is to estimate the values of the hyperparameters as well as the posterior probability distribution of the latent variables. Within the conventional variational inference framework, the objective function that needs to be maximized is

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

For most of the non-Gaussian mixture models (*e.g.*, the beta mixture model [7], the Dirichlet mixture model [6], the beta-Liouville mixture model [34], the inverted Dirichlet mixture model [17]), the term $\mathrm{E}_{q(\Theta)}[\ln p(\mathcal{X}, \Theta)]$ is analytically intractable such that the lower bound $\mathcal{L}(q)$ cannot be maximized directly by a closed-form solution. Therefore, the EVI method [6], [7], [41] was proposed to overcome the aforementioned problem. With an auxiliary function $\tilde{p}(\mathcal{X}, \Theta)$ that satisfies

$$\mathrm{E}_{q(\Theta)}[\ln p(\mathcal{X}, \Theta)] \geq \mathrm{E}_{q(\Theta)}[\ln \tilde{p}(\mathcal{X}, \Theta)] \tag{19}$$

and substituting (19) into (18), we can still reach the maximum value of $\mathcal{L}(q)$ at some given points by maximizing a lower bound of $\tilde{\mathcal{L}}(q)$

$$\mathcal{L}(q) \geq \tilde{\mathcal{L}}(q) = \mathrm{E}_{q(\Theta)}[\ln \tilde{p}(\mathcal{X}, \Theta)] - \mathrm{E}_{q(\Theta)}[\ln q(\Theta)]. \tag{20}$$

If $\tilde{p}(\mathcal{X}, \Theta)$ is properly selected, an analytically tractable solution can be obtained. In order to properly formulate the variational posterior $q(\Theta)$, we truncate the stick-breaking representation for the InIDMM at a value M as

$$\lambda_M = 1, \quad \pi_m = 0 \quad \text{when } m > M, \quad \text{and} \sum_{m=1}^{M} \pi_m = 1. \tag{21}$$

Note that the model is still a full DP mixture. The truncation level M is not a part of our prior infinite mixture model, it is only a variational parameter for pursuing an approximation to the posterior, which can be freely initialized and automatically optimized without yielding overfitting during the learning process. Additionally, we make use of the following factorized variational distribution to approximate $p(\Theta|\mathcal{X})$ as

$$q(\Theta) = \prod_{m=1}^{M} q(\lambda_m) q(\varphi_m) \prod_{n=1}^{N} q(z_{nm}) \prod_{d=1}^{D+1} q(\alpha_{md}), \tag{22}$$

where the variables in the posterior distribution are assumed to be mutually independent (as illustrated by the graphical model in Fig. 1). This is the only assumption we introduced to the posterior distribution. No other restrictions are imposed over the mathematical forms of the individual factor distributions [2].

Applying the full factorization formulation and the truncated stick-breaking representation for the proposed model, we can solve the variational learning by maximizing the lower bound $\tilde{\mathcal{L}}(q)$ shown in (20). The optimal solution in this case is given by

$$\ln q_s(\Theta_s) = \langle \ln \tilde{p}(\mathcal{X}, \Theta) \rangle_{j \neq s} + \text{Con.}, \tag{23}$$

where $\langle \cdot \rangle_{j \neq s}$ refers to the expectation with respect to all the distributions $q_j(\Theta_j)$ except for variable s. In addition, any term that does not include Θ_s are absorbed into the additive constant "Con." [2], [41]. In the variational inference, all factors $q_s(\Theta_s)$ need to be suitably initiated, then each factor is updated in turn with a revised value obtained by (23) using the current values of all the other factors. Convergence is theoretically guaranteed since the lower bound is a convex with respect to each factor $q_s(\Theta_s)$ [2], [6].

B. *EVI for the Optimal Posterior Distributions*

According to the principles of EVI, the expectation of the logarithm of the joint distribution, given the joint posterior distributions of the parameters, can be expressed as

$$\begin{aligned}
&\langle \ln p(\mathcal{X}, \Theta) \rangle \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} \langle z_{nm} \rangle \Bigg[\mathcal{R}_m + \sum_{d=1}^{D} (\langle \alpha_{md} \rangle - 1) \ln x_{nd} \\
&\quad - \sum_{d=1}^{D+1} \langle \alpha_{md} \rangle (1 + \sum_{d=1}^{D} x_{nd}) + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1 - \lambda_j) \rangle \Bigg] \\
&\quad + \sum_{m=1}^{M} [\langle \ln \varphi_m \rangle + (\langle \varphi_m \rangle - 1) \langle \ln(1 - \lambda_m) \rangle] \\
&\quad + \sum_{m=1}^{M} \sum_{d=1}^{D+1} [(u_{md} - 1) \langle \ln \alpha_{md} \rangle - v_{md} \langle \alpha_{md} \rangle] \\
&\quad + \sum_{m=1}^{M} [(s_m - 1) \langle \ln \varphi_m \rangle - t_m \langle \varphi_m \rangle] + \text{Con.},
\end{aligned} \tag{24}$$

where $\mathcal{R}_m = \left\langle \ln \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_{md})}{\prod_{d=1}^{D+1} \Gamma(\alpha_{md})} \right\rangle$.

With the mathematical expression in (24), an analytically tractable solution is not feasible, which is due to the fact that $\mathcal{R}_m$ cannot be explicitly calculated (although it can be simulated by some numerical sampling methods). In order to apply (23) to explicitly calculate the optimal posterior distributions and with the principles of the EVI framework, it is required to introduce an auxiliary function $\tilde{\mathcal{R}}_m$ such that $\mathcal{R}_m \geq \tilde{\mathcal{R}}_m$. According to [6, Eq. 25], we can select $\tilde{\mathcal{R}}_m$ as

$$\begin{aligned}
\tilde{\mathcal{R}}_m = \ln \frac{\Gamma(\sum_{d=1}^{D+1} \langle \alpha_{md} \rangle)}{\prod_{d=1}^{D+1} \Gamma(\langle \alpha_{md} \rangle)} + \sum_{d=1}^{D+1} \left[\Psi(\sum_{k=1}^{D+1} \langle \alpha_{md} \rangle) - \Psi(\langle \alpha_{md} \rangle) \right] \\
\times [\langle \ln \alpha_{md} \rangle - \ln \langle \alpha_{md} \rangle] \langle \alpha_{md} \rangle,
\end{aligned} \tag{25}$$

where $\Psi(\cdot)$ is the digamma function defined as $\Psi(a) = \partial \ln \Gamma(a) / \partial a$.

Substituting (25) into (24), a lower bound to $\langle \ln p(\mathcal{X}, \Theta) \rangle$ can be obtained as

$$\begin{aligned}
&\langle \ln \tilde{p}(\mathcal{X}, \Theta) \rangle \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} \langle z_{nm} \rangle \Bigg[\tilde{\mathcal{R}}_m + \sum_{d=1}^{D} (\langle \alpha_{md} \rangle - 1) \ln x_{nd} \\
&\quad - \sum_{d=1}^{D+1} \langle \alpha_{md} \rangle (1 + \sum_{d=1}^{D} x_{nd}) + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln(1 - \lambda_j) \rangle \Bigg] \\
&\quad + \sum_{m=1}^{M} [\langle \ln \varphi_m \rangle + (\langle \varphi_m \rangle - 1) \langle \ln(1 - \lambda_m) \rangle] \\
&\quad + \sum_{m=1}^{M} \sum_{d=1}^{D+1} [(u_{md} - 1) \langle \ln \alpha_{md} \rangle - v_{md} \langle \alpha_{md} \rangle] \\
&\quad + \sum_{m=1}^{M} [(s_m - 1) \langle \ln \varphi_m \rangle - t_m \langle \varphi_m \rangle] + \text{Con.}.
\end{aligned} \tag{26}$$

With (23), we can get analytically tractable solutions for optimally estimating the posterior distributions of $\mathcal{Z}$, $\vec{\lambda}$, $\vec{\varphi}$, and Λ. We now consider each of these in more detail: *1) The posterior distribution of $q(\mathcal{Z})$*

As any term that is independent of z_{nm} can be absorbed into the additive constant, we have

$$\ln q^*(z_{nm}) = \text{Con.} + z_{nm}\Bigg[\widetilde{\mathcal{R}}_m + \langle \ln \lambda_m\rangle + \sum_{j=1}^{m-1}\langle \ln(1-\lambda_j)\rangle + \sum_{d=1}^{D}(\langle \alpha_{md}\rangle - 1)\ln x_{nd} + \sum_{d=1}^{D+1}\langle \alpha_{md}\rangle \ln(1+\sum_{d=1}^{D} x_{nd})\Bigg], \tag{27}$$

which has same logarithmic form of the prior distribution (*i.e.*, the categorial distribution). Therefore, we can write $\ln q^*(\mathcal{Z})$ as

$$\ln q^*(\mathcal{Z}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\ln \rho_{nm} + \text{Con.} \tag{28}$$

with the definition that

$$\ln \rho_{nm} = \langle \ln \lambda_m\rangle + \sum_{j=1}^{m-1}\langle \ln(1-\lambda_j)\rangle + \tilde{\mathcal{R}}_m + \sum_{d=1}^{D}(\langle \alpha_{md}\rangle - 1)\ln x_{nd} - \sum_{d=1}^{D+1}\langle \alpha_{md}\rangle(1+\sum_{d=1}^{D} x_{nd}). \tag{29}$$

Recalling that $z_{nm} \in (0,1)$ and $\sum_{m=1}^{M} z_{nm} = 1$, we define

$$r_{nm} = \frac{\rho_{nm}}{\sum_{m=1}^{M}\rho_{nm}}. \tag{30}$$

Taking the exponential of both sides of (28), we have

$$q^*(\mathcal{Z}) = \prod_{n=1}^{N}\prod_{m=1}^{M} r_{nm}^{z_{nm}}, \tag{31}$$

which is the optimal posterior distribution of $\mathcal{Z}$.

The posterior mean $\langle z_{nm}\rangle$ can be calculated as $\langle z_{nm}\rangle = r_{nm}$. Actually, the quantities $\{r_{nm}\}$ are playing a similar role as the responsibilities in the conventional EM [51] algorithm.

In the following parts, we show only the optimal solutions to $\vec{\lambda}$, $\vec{\varphi}$, and Λ, respectively. The derivation details can be found in the appendix.

2) The posterior distribution of $q(\vec{\lambda})$

The optimal solution to the posterior distribution of $\vec{\lambda}$ is characterized as

$$q(\vec{\lambda}) = \prod_{m=1}^{M}\text{Beta}(\lambda_m; g_m^*, h_m^*), \tag{32}$$

where the hyperparameters s_m^* and q_m^* are

$$g_m^* = 1 + \sum_{n=1}^{N}\langle z_{nm}\rangle, \quad h_m^* = \langle \varphi_m\rangle + \sum_{n=1}^{N}\sum_{j=m+1}^{M}\langle z_{nj}\rangle. \tag{33}$$

3) The posterior distribution of $q(\vec{\varphi})$

The optimal solution to the posterior distribution of $\vec{\varphi}$ is

$$q^*(\vec{\varphi}) = \prod_{m=1}^{M}\text{Gam}(\varphi_m; s_m^*, t_m^*), \tag{34}$$

where the optimal solutions to the hyperparamters s_m^* and t_m^* are

$$s_m^* = 1 + s_m^0, \quad t_m^* = t_m^0 - \langle \ln(1-\lambda_m)\rangle, \tag{35}$$

where s_m^0 and t_m^0 denote the hyperparameters initialized in the prior distribution, respectively.

4) The posterior distribution of $q(\Lambda)$

The optimal approximation to the posterior distribution of Λ is

$$q^*(\Lambda) = \prod_{m=1}^{M}\prod_{d=1}^{D+1}\text{Gam}(\alpha_{md}; u_{md}^*, v_{md}^*), \tag{36}$$

where the optimal solutions to the hyperparameters u_{md}^* and v_{md}^* are given by

$$u_{md}^* = u_{md}^0 + \sum_{n=1}^{N}\langle z_{nm}\rangle\left[\Psi(\sum_{k=1}^{K+1}\langle \alpha_{mk}\rangle) - \Psi(\langle \alpha_{md}\rangle)\right]\langle \alpha_{md}\rangle \tag{37}$$

and

$$v_{md}^* = v_{md}^0 - \sum_{n=1}^{N}\langle z_{nm}\rangle\left[\ln x_{nd} - \ln(1+\sum_{d=1}^{D} x_{nd})\right]. \tag{38}$$

In the above equations, u_{md}^0 and v_{md}^0 are the hyperparameters in the prior distribution and we set $x_{n,D+1} = 1$. The following expectations are needed to calculate the aforementioned update equations:

$$\begin{aligned}\langle \ln(1-\lambda_m)\rangle =&\Psi(h_m^*) - \Psi(g_m^* + h_m^*),\\ \langle \ln \lambda_m\rangle =&\Psi(g_m^*) - \Psi(g_m^* + t_m^*),\\ \langle \ln \alpha_{md}\rangle =&\Psi(u_{md}^*) - \ln v_{md}^*,\\ \langle \varphi_m\rangle = \frac{s_m^*}{t_m^*},\ &\langle \alpha_{md}\rangle = \frac{u_{md}^*}{v_{md}^*}.\end{aligned} \tag{39}$$

C. *Full Variational Learning Algorithm*

As can be observed from the above updating process, the optimal solutions for the posterior distributions are dependent on the moments evaluated with respect to the posterior distributions of the other variables. Thus, the variational update equations are mutually coupled. In order to obtain optimal posterior distributions for all the variables, iterative updates are required until convergence. With the obtained posterior distributions, it is straightforward to calculate the lower bound $\tilde{\mathcal{L}}(q)$

$$\begin{aligned}\tilde{\mathcal{L}}(q) =& \int q(\Theta)\ln\frac{\tilde{p}(\Theta,\mathcal{X})}{q(\Theta)}d\Theta\\ =&\langle \ln \tilde{p}(\mathcal{X},\Theta)\rangle - \langle \ln q(\Theta)\rangle\\ =&\langle \ln \tilde{p}(\mathcal{X},\Theta)\rangle - \langle \ln q(\mathcal{Z})\rangle - \langle \ln q(\vec{\lambda})\rangle\\ &- \langle \ln q(\vec{\varphi})\rangle - \langle \ln q(\Lambda)\rangle,\end{aligned} \tag{40}$$

which is helpful in monitoring the convergence. In (40), each term with expectation (*i.e.*, $\langle\cdot\rangle$) is evaluated with respect to all the variables in its argument as

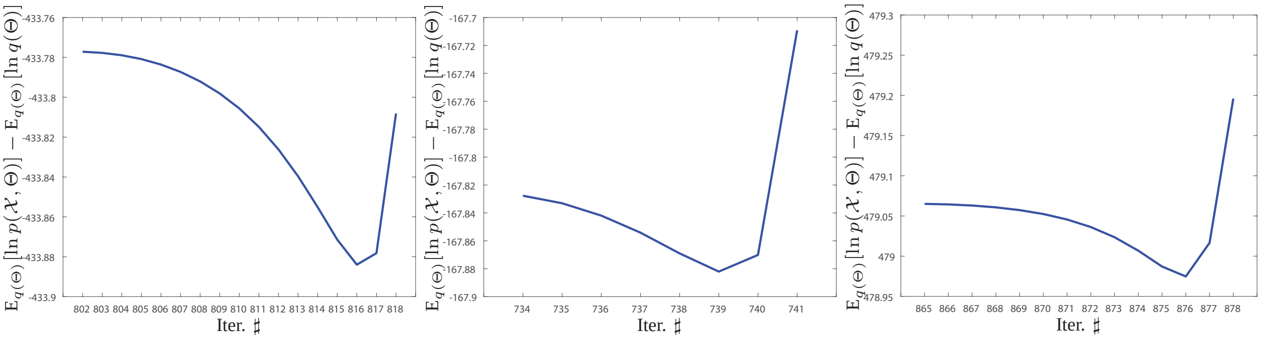
$$\langle \ln q(\mathcal{Z})\rangle = r_{nm}\ln r_{nm}, \tag{41}$$

$$\begin{aligned}\langle \ln q(\vec{\lambda})\rangle = \sum_{m=1}^{M}&\left[\ln\Gamma(g_m^* + h_m^*) - \ln\Gamma(g_m^*) - \ln\Gamma(h_m^*)\right.\\ &\left.+(g_m^* - 1)\langle \ln \lambda_m\rangle + (h_m^* - 1)\langle \ln(1-\lambda_m)\rangle\right],\end{aligned} \tag{42}$$

$$\begin{aligned}\langle \ln q(\vec{\varphi})\rangle = \sum_{m=1}^{M}&\left[s_m^*\ln t_m^* - \ln\Gamma(s_m^*)\right.\\ &\left.+(s_m^* - 1)\langle \ln \varphi_m\rangle - t_m^*\bar{\varphi}_m\right],\end{aligned} \tag{43}$$

and

$$\begin{aligned}\langle \ln q(\vec{\alpha})\rangle = \sum_{m=1}^{M}\sum_{d=1}^{D+1}&\left[u_{md}^*\ln v_{md}^* - \ln\Gamma(u_m^*)\right.\\ &\left.+(u_m^* - 1)\langle \ln \alpha_{md}\rangle - v_{md}^*\bar{\alpha}_{md}\right].\end{aligned} \tag{44}$$



(a) Model A

(b) Model B

(c) Model C

Fig. 3: Observations of the objective function's oscillations during iterations. This non-convergence indicates that the MLB approximation-based method cannot theoretically guarantee convergence. The model settings are the same as Tab. I.

Algorithm 1 Algorithm for EVI-based Bayesian InIDMM

```
Set the initial truncation level $M$ and the initial values for hyperparameters $s_m^0$, $t_m^0$, $u_{md}^0$, and $v_{md}^0$
Initialize the values of $r_{nm}$ by $K$-means algorithm.
repeat
    Calculate the expectations in (39).
    Update the posterior distributions for each variable by (33), (35), (37) and (38).
until Stop criterion is reached.
For all $m$, calculate $\langle \lambda_m \rangle = s_m^*/(s_m^* + t_m^*)$ and substitute it back into (11) to get the estimated values of the mixing coefficients $\widehat{\pi}_m$.
Determine the optimum number of components $M$ by eliminating the components with mixing weights smaller than $10^{-5}$.[3]
Renormalize $\{\widehat{\pi}_m\}$ to have a unit $l_1$ norm.
Calculate $\widehat{\alpha}_{md} = u_{md}^*/v_{md}^*$ for all $m$ and $d$.
```

Additionally, $\langle \ln \tilde{p}(\mathcal{X},\Theta)\rangle$ is given in (26) .

The algorithm of the proposed EVI-based Bayesian estimation of InIDMM is summarized in Algorithm 1.

IV. Experimental Results and Discussions

In this section, both synthesized data and real data are utilized to demonstrate the performance of the proposed algorithm for InIDMM. In the initialization stage of all the experiments, the truncation level M is set to 15 and the hyperparameters of the gamma prior distributions are chosen as $u_0 = s_0 = 1$ and $v_0 = t_0 = 0.005$, which provide non-informative prior distributions. Note that these specific choices were based on our experiments and were found convenient and effective in our case. We take the posterior means as point estimates to the parameters in an InIDMM.

A. Synthesized Data Evaluation

As shown in the previous studies for EVI-based Bayesian estimation [5], [6], the SLB approximation can guarantee the convergence while the MLB approximation cannot. We use the synthesized data evaluation to compare the Bayesian InIDMM using the SLB approximation (proposed in this paper and denoted as InIDMM$_{\text{SLB}}$) with the Bayesian InIDMM using the MLB approximation (proposed in [46] and denoted as InIDMM$_{\text{MLB}}$). Three models (see Tab. I for details) were selected to generate the synthesized datasets.

1) Model Selection: One advantage of DP process mixture model is to decide the number of mixture components automatically, based on the training data. Following the instructions in [52] and for a first check, we ran the proposed EVI-based method for InIDMM$_{\text{SLB}}$. The optimization procedure is carried out without component elimination (*i.e.*, a fixed number of components, M, is chosen and the mixing coefficients are fixed during iteration. The initial value of the mixing coefficients were obtained from plain EM estimation.) Under this setting, the variational lower-bound can be treated as a model selection score and the effect of the number of the mixture components is demonstrated. With synthesized data generated from the aforementioned three models, we plotted the relation between the variaional lower-bounds and the number of mixture components in Fig. 4.

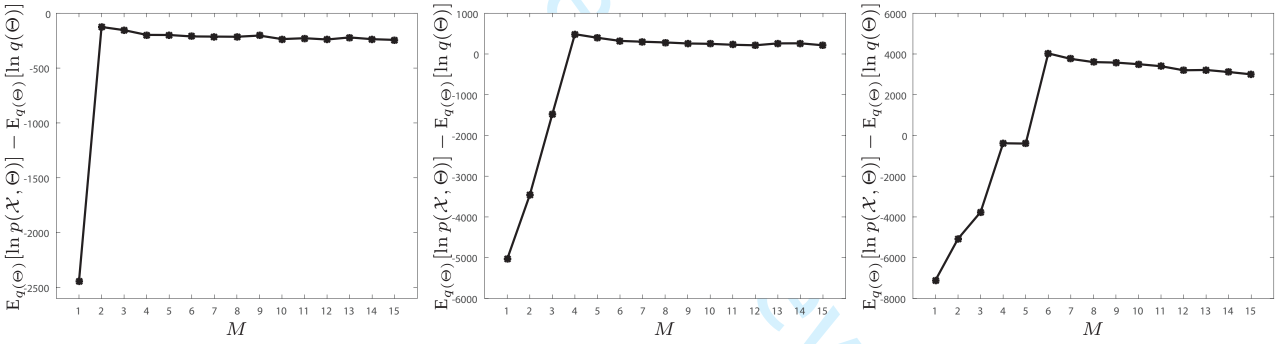
2) Observations of Oscillations: We ran the InIDMM$_{\text{MLB}}$ algorithm and monitored the value of the variational objective function during each iteration. It can be observed that the variational objective function was not always increasing in Bayesian estimation with the InIDMM$_{\text{MLB}}$. Figure 3 illustrates the decreasing values during iterations. On the other hand, the variational objective function obtained with the InIDMM$_{\text{SLB}}$ algorithm was always increasing until convergence, as the SLB approximation insures the convergency theoretically. The observations of oscillations demonstrate that the convergence with MLB approximation cannot be guaranteed. The original variational object function was numerically calculated by employing sampling method. In order to monitor the parameter estimation process of InIDMM$_{\text{SLB}}$, we show the value of the variational objective function during iterations in Fig. 5. It can be observe that the variational objective function obtained by InIDMM$_{\text{SLB}}$ increases during iterations and in most cases it increases very fast.

3) Quantitative Comparisons: Next, we compare the InIDMM$_{\text{SLB}}$ with the InIDMM$_{\text{MLB}}$ quantitatively. With a known IDMM, 2000 samples were generated. The InIDMM$_{\text{SLB}}$ and the InIDMM$_{\text{MLB}}$ were applied to estimate the posterior distributions of the model, respectively. In Tab. I,

[3] When a mixing coefficient is small enough, it converges to 0 faster. Therefore, we can remove components with very small value (less than a threshold). This choice (empirically choosing a threshold) is purely for the convenience of easy implementation. Similar strategy is also widely used applied in many other sticking-break process-based DP mixture models, *e.g.*, [34], [46].

TABLE I: Comparisons of true and estimated models.

True Model	Model A $\pi_1 = 0.5 \ \ , \vec{\alpha}_1 = [16 \ 8 \ 6 \ 2]^{\mathrm{T}}$ $\pi_2 = 0.5 \ \ , \vec{\alpha}_2 = [8 \ 12 \ 15 \ 18]^{\mathrm{T}}$	Model B $\pi_1 = 0.25 \ \ , \vec{\alpha}_1 = [12 \ 36 \ 14 \ 18 \ 55 \ 16]^{\mathrm{T}}$ $\pi_2 = 0.25 \ \ , \vec{\alpha}_2 = [32 \ 48 \ 25 \ 12 \ 36 \ 48]^{\mathrm{T}}$ $\pi_3 = 0.25 \ \ , \vec{\alpha}_3 = [25 \ 10 \ 18 \ 10 \ 36 \ 48]^{\mathrm{T}}$ $\pi_4 = 0.25 \ \ , \vec{\alpha}_4 = [6 \ 28 \ 16 \ 32 \ 12 \ 24]^{\mathrm{T}}$
InIDMM$_{\mathrm{SLB}}$	$\widehat{\pi}_1 = 0.502 \ \ , \widehat{\vec{\alpha}}_1 = [16.96 \ 8.58 \ 6.39 \ 12.49]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.498 \ \ , \widehat{\vec{\alpha}}_2 = [8.20 \ 12.16 \ 15.49 \ 18.34]^{\mathrm{T}}$	$\widehat{\pi}_1 = 0.251 \ \ , \widehat{\vec{\alpha}}_1 = [12.26 \ 36.59 \ 14.30 \ 18.19 \ 56.36 \ 16.25]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.249 \ \ , \widehat{\vec{\alpha}}_2 = [33.37 \ 49.92 \ 25.85 \ 12.80 \ 37.00 \ 49.79]^{\mathrm{T}}$ $\widehat{\pi}_3 = 0.252 \ \ , \widehat{\vec{\alpha}}_3 = [25.72 \ 10.32 \ 18.09 \ 10.09 \ 37.27 \ 49.58]^{\mathrm{T}}$ $\widehat{\pi}_4 = 0.248 \ \ , \widehat{\vec{\alpha}}_4 = [6.14 \ 28.94 \ 16.72 \ 33.46 \ 12.32 \ 25.20]^{\mathrm{T}}$
InIDMM$_{\mathrm{MLB}}$	$\widehat{\pi}_1 = 0.508 \ \ , \widehat{\vec{\alpha}}_1 = [15.20 \ 7.71 \ 5.90 \ 11.64]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.492 \ \ , \widehat{\vec{\alpha}}_2 = [9.21 \ 13.76 \ 17.13 \ 21.10]^{\mathrm{T}}$	$\widehat{\pi}_1 = 0.249 \ \ , \widehat{\vec{\alpha}}_1 = [12.18 \ 37.82 \ 14.56 \ 18.85 \ 57.32 \ 16.44]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.249 \ \ , \widehat{\vec{\alpha}}_2 = [33.71 \ 51.10 \ 26.92 \ 12.89 \ 38.66 \ 51.73]^{\mathrm{T}}$ $\widehat{\pi}_3 = 0.250 \ \ , \widehat{\vec{\alpha}}_3 = [24.94 \ 9.90 \ 18.07 \ 10.04 \ 36.10 \ 48.25]^{\mathrm{T}}$ $\widehat{\pi}_4 = 0.252 \ \ , \widehat{\vec{\alpha}}_4 = [5.82 \ 27.43 \ 15.77 \ 31.14 \ 11.82 \ 23.58]^{\mathrm{T}}$
True Model	Model C $\pi_1 = 0.2 \ \ , \vec{\alpha}_1 = [12 \ 21 \ 36 \ 18 \ 32 \ 65 \ 76]^{\mathrm{T}}$ $\pi_2 = 0.2 \ \ , \vec{\alpha}_2 = [28 \ 42 \ 21 \ 8 \ 54 \ 21 \ 48]^{\mathrm{T}}$ $\pi_3 = 0.2 \ \ , \vec{\alpha}_3 = [32 \ 12 \ 7 \ 35 \ 13 \ 32 \ 18]^{\mathrm{T}}$ $\pi_4 = 0.2 \ \ , \vec{\alpha}_4 = [62 \ 44 \ 31 \ 65 \ 72 \ 15 \ 44]^{\mathrm{T}}$ $\pi_5 = 0.2 \ \ , \vec{\alpha}_5 = [53 \ 12 \ 18 \ 44 \ 65 \ 33 \ 52]^{\mathrm{T}}$	
InIDMM$_{\mathrm{SLB}}$	$\widehat{\pi}_1 = 0.201 \ \ , \widehat{\vec{\alpha}}_1 = [12.08 \ 20.89 \ 36.25 \ 18.28 \ 32.69 \ 65.72 \ 76.70]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.199 \ \ , \widehat{\vec{\alpha}}_2 = [29.12 \ 43.43 \ 21.41 \ 8.33 \ 56.11 \ 21.74 \ 49.20]^{\mathrm{T}}$ $\widehat{\pi}_3 = 0.200 \ \ , \widehat{\vec{\alpha}}_3 = [31.57 \ 11.89 \ 6.99 \ 34.70 \ 12.90 \ 31.85 \ 17.89]^{\mathrm{T}}$ $\widehat{\pi}_4 = 0.201 \ \ , \widehat{\vec{\alpha}}_4 = [59.83 \ 42.55 \ 29.89 \ 61.98 \ 67.68 \ 14.11 \ 42.46]^{\mathrm{T}}$ $\widehat{\pi}_5 = 0.199 \ \ , \widehat{\vec{\alpha}}_5 = [58.00 \ 12.8 \ 20.02 \ 47.70 \ 71.08 \ 36.57 \ 57.66]^{\mathrm{T}}$	
InIDMM$_{\mathrm{MLB}}$	$\widehat{\pi}_1 = 0.200 \ \ , \widehat{\vec{\alpha}}_1 = [12.56 \ 21.50 \ 37.69 \ 19.00 \ 33.06 \ 68.04 \ 79.64]^{\mathrm{T}}$ $\widehat{\pi}_2 = 0.200 \ \ , \widehat{\vec{\alpha}}_2 = [28.26 \ 43.02 \ 20.85 \ 8.14 \ 55.36 \ 21.21 \ 49.17]^{\mathrm{T}}$ $\widehat{\pi}_3 = 0.199 \ \ , \widehat{\vec{\alpha}}_3 = [32.17 \ 12.19 \ 7.13 \ 35.66 \ 13.01 \ 32.54 \ 17.84]^{\mathrm{T}}$ $\widehat{\pi}_4 = 0.199 \ \ , \widehat{\vec{\alpha}}_4 = [63.61 \ 45.48 \ 32.00 \ 66.63 \ 74.31 \ 15.21 \ 45.45]^{\mathrm{T}}$ $\widehat{\pi}_5 = 0.202 \ \ , \widehat{\vec{\alpha}}_5 = [52.12 \ 11.83 \ 18.34 \ 43.77 \ 64.80 \ 32.53 \ 51.48]^{\mathrm{T}}$	



(a) Model A (b) Model B (c) Model C

Fig. 4: Effect of the number of mixture components.

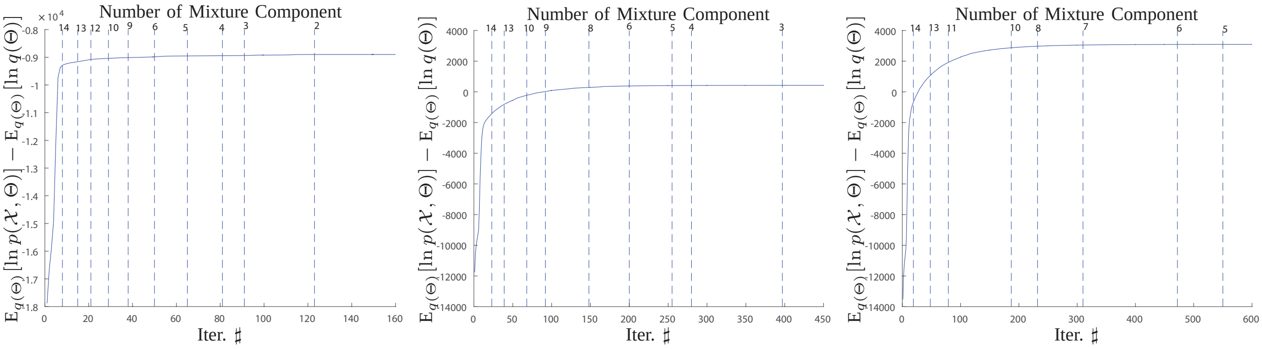
TABLE II: Comparisons of objective function values and runtime for InIDMM with SLB and MLB.

Model & Method	Model A		Model B		Model C	
	InIDMM$_{\mathrm{SLB}}$	InIDMM$_{\mathrm{MLB}}$	InIDMM$_{\mathrm{SLB}}$	InIDMM$_{\mathrm{MLB}}$	InIDMM$_{\mathrm{SLB}}$	InIDMM$_{\mathrm{MLB}}$
Obj. Func. Val.	$\mathbf{-1.86 \times 10^3}$	-1.90×10^3	$\mathbf{0.42 \times 10^3}$	0.32×10^3	$\mathbf{3.05 \times 10^3}$	2.99×10^3
p-values	0.046		6.48×10^{-4}		0.016	
$\mathrm{KL}(p(\mathcal{X}\|\Theta)\\|p(\mathcal{X}\|\widehat{\Theta}))$	$\mathbf{3.35 \times 10^{-3}}$	6.97×10^{-3}	$\mathbf{2.80 \times 10^{-3}}$	8.07×10^{-3}	$\mathbf{2.93 \times 10^{-3}}$	6.24×10^{-3}
p-values	1.46×10^{-11}		6.93×10^{-15}		2.08×10^{-7}	
Runtime (in s)†	**2.06**	2.26	**3.06**	3.61	**2.84**	3.07

† On a ThinkCentre® computer with Intel® Core™ i5 – 4590 CPU 8G.

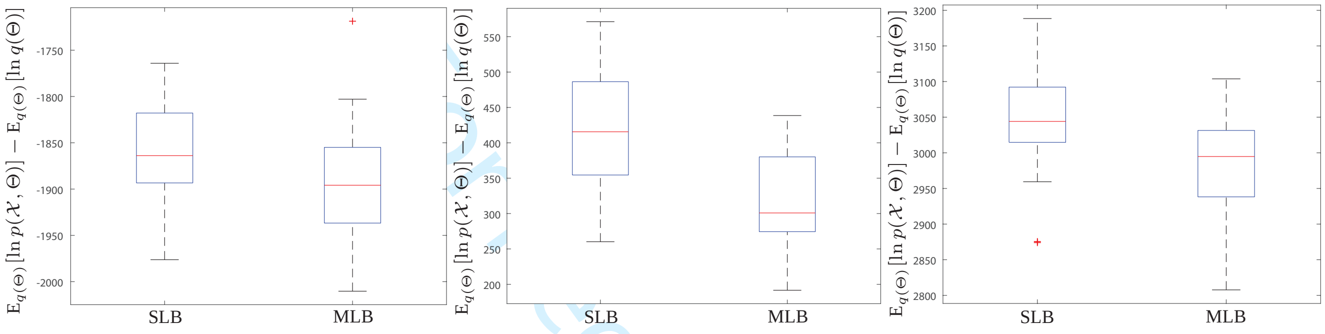
we list the estimated parameters by taking the posterior means. It can be observed that, both the InIDMM$_{\mathrm{SLB}}$ and the InIDMM$_{\mathrm{MLB}}$ can carry out the estimation properly. However, with 20 repeats of the aforementioned "data generation-model estimation" procedure and calculating the variational objective function with sampling method, superior performance of the InIDMM$_{\mathrm{SLB}}$ over the InIDMM$_{\mathrm{MLB}}$ can be observed from Tab. II. The mean values of the objective function obtained by InIDMM$_{\mathrm{SLB}}$ are larger than those obtained by the InIDMM$_{\mathrm{SLB}}$ while the computational cost (measured in seconds) required by the InIDMM$_{\mathrm{SLB}}$ are smaller than those required by the InIDMM$_{\mathrm{MLB}}$. Moreover, smaller KL divergences[4] of the estimated models from the corresponding true models also verify that the InIDMM$_{\mathrm{SLB}}$ yields better estimates than the InIDMM$_{\mathrm{MLB}}$. In order to examine if the differences between the InIDMM$_{\mathrm{SLB}}$ and the InIDMM$_{\mathrm{MLB}}$ are statistically significant, we conducted the student's t-test with the null-

[4] Here, the KL divergence is calculated as $\mathrm{KL}(p(\mathcal{X}|\Theta)\|p(\mathcal{X}|\widehat{\Theta}))$ by sampling method. $\widehat{\Theta}$ denotes the point estimate of the parameters from the posterior distribution.



(a) Model A (b) Model B (c) Model C

Fig. 5: Illustration of the variational objective function's values obtained by SLB against the number of iterations.



(a) Model A (b) Model B (c) Model C

Fig. 6: Boxplots for comparisons of the objective function values' distributions obtained by SLB and MLB with different models. The model settings are the same as those in Tab. I. The central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles. The outliers are marked individually.

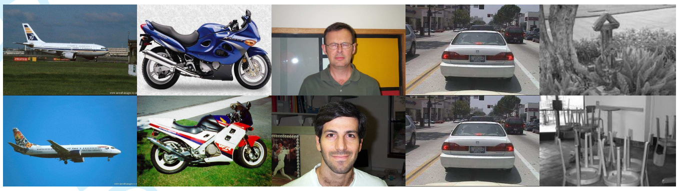
TABLE III: Comparisons of image categorization accuracies (in %) obtained with different models. The standard deviations are in the brackets. The p-values of the student's t-test with the null-hypothesis that InIDMM$_{\text{SLB}}$ and the referring method have equal means but unknown variances are listed.

	InIDMM$_{\text{SLB}}$	IDMM$_{\text{SLB}}$	InIDMM$_{\text{MCMC}}$	InGMM	SVM
Caltech-4	**93.49**	89.27	90.21	83.92	92.72
	(1.05)	(0.84)	(0.73)	(0.72)	(0.82)
p-value	N/A	1.01×10^{-8}	1.91×10^{-7}	4.55×10^{-15}	0.085
ETH-80	**75.49**	72.88	73.05	68.88	72.47
	(0.75)	(1.46)	(0.78)	(0.74)	(0.70)
p-value	N/A	8.69×10^{-5}	1.17×10^{-6}	1.60×10^{-13}	2.49×10^{-8}

hypothesis that the results obtained by these two methods have equal means and equal but unknown variances. All the p-values of in Tab. II are smaller than the significant level 0.1, which indicates that the superiority of the InIDMM$_{\text{SLB}}$ over the InIDMM$_{\text{MLB}}$ is statistically significant. The distributions of the objective function values are shown by the boxplots in Fig. 6.

B. Real Data Evaluation

In the real data evaluations, the proposed InIDMM$_{\text{SLB}}$ has been applied for the task of image categorization and object detection. The referred methods for comparisons are the IDMM$_{\text{SLB}}$ [53], the Markov Chain Monte Carlo-based numerical model estimation (InIDMM$_{\text{MCMC}}$, numerical simulation of the posterior distributions) [54], the Dirichlet process Gaussian mixture model (InGMM, another commonly used statistical model) [55], and the support vector machine (SVM)-based classifier (discriminant method, implemented with LIBSVM toolbox [56]).



(a) Airplane (b) Motorbike (c) Face (d) Car (e) Background

Fig. 7: Sample images from the Caltech-4 dataset.

1) Datasets: The evaluations were conducted based on two well-known datasets. The first dataset is the Caltech-4 dataset [5]. It is a composite of four different categories. They are 1074 images of airplanes from the side, 526 images of cars from the rear, 826 images of motorbikes from the side, and 450 frontal face images from about 27 unique persons. Example images from these four categories are shown in Fig. 7(a)-7(d). The second dataset is the ETH-80 dataset [6] that consists of eight categories: apple, car, cup, dog, pear, tomato, horse, and cow. Each category has 410 images which are cropped, so that they contain only the object in the center. Examples of images from each category in the ETH-80 dataset are shown in Fig. 9. Our experiments were evaluated on the these two commonly used public datasets for the purpose of validating the effectiveness of the proposed method.

2) Descriptor Extraction: In recent years, many excellent global and local descriptors have been proposed for the purpose of image categorization and object detection. For

[5] http://www.vision.caltech.edu/archive.html

[6] http://www.d2.mpi-inf.mpg.de/Datasets/ETH80

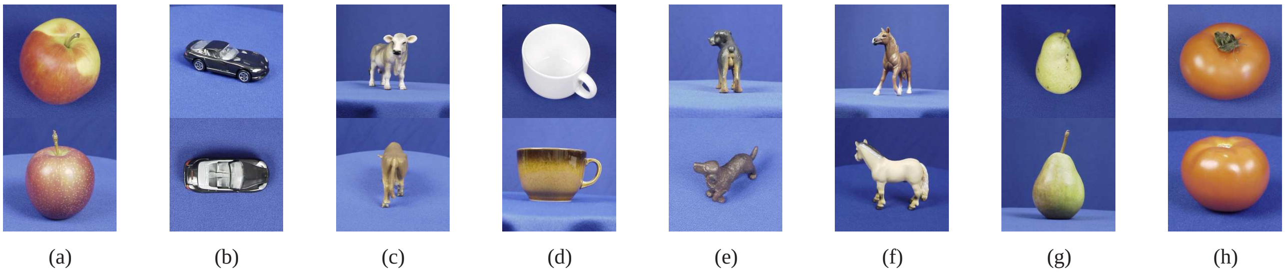
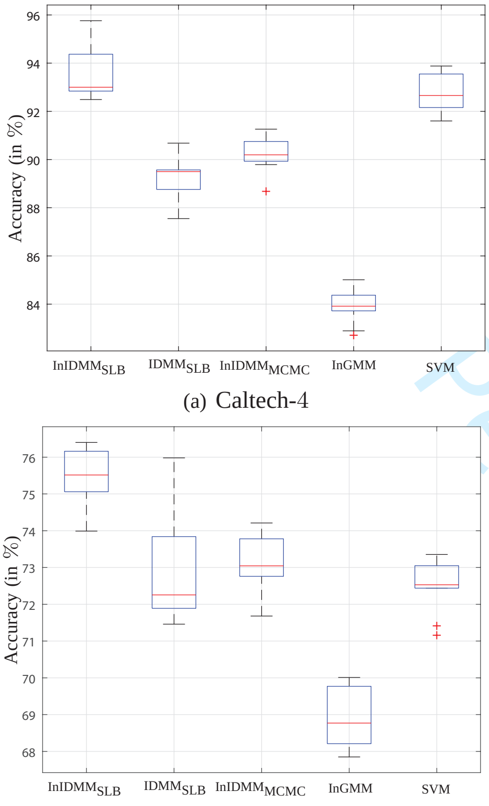


Fig. 9: Sample images from ETH-80 dataset. (a) Apple. (b) Car. (c) Cow. (d) Cup. (e) Dog. (f) Horse. (g) Pear. (h) Tomato.



(a) Caltech-4

(b) ETH-80

Fig. 8: Boxplots for comparisons of the categorization accuracies' distributions for the Caltech-4 and the ETH-80 datasets. The central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles. The outliers are marked individually.

TABLE IV: Comparisons of object detections accuracies (in $\%$) on Caltech-4 dataset. The standard deviations are in the brackets. The p-values of the student's t-test with the null-hypothesis that InIDMM$_{\text{SLB}}$ and the referring method have equal means but unknown variances are listed.

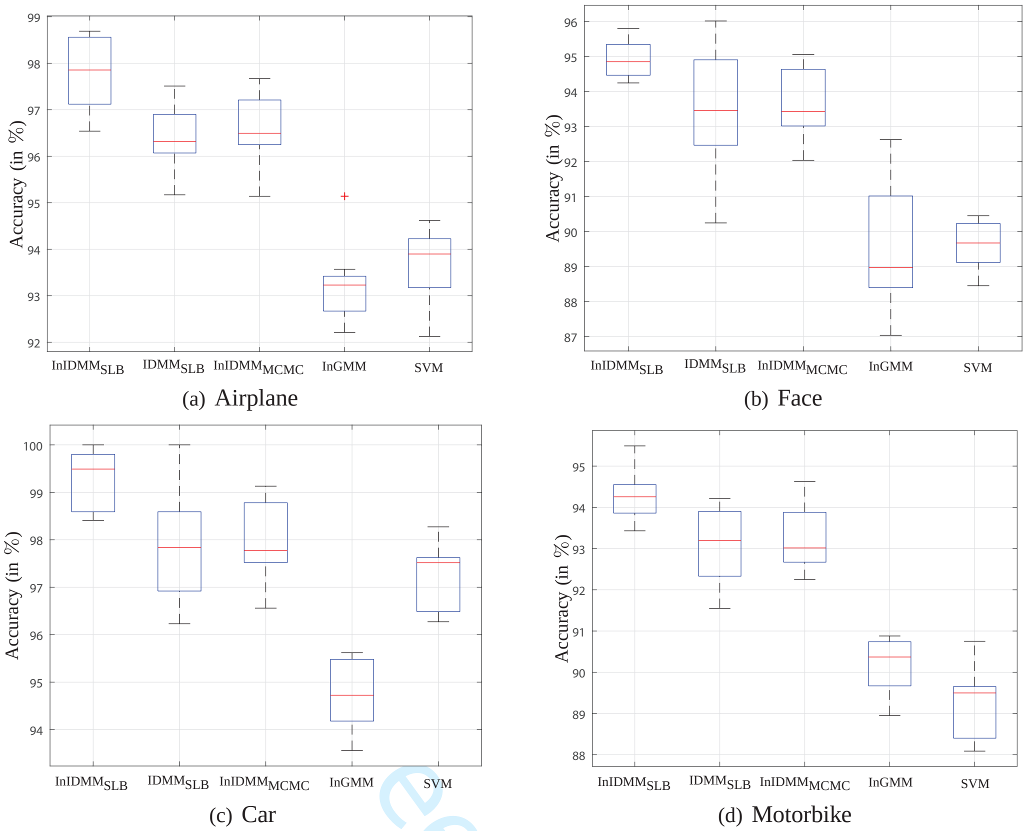
	InIDMM$_{\text{SLB}}$	IDMM$_{\text{SLB}}$	InIDMM$_{\text{MCMC}}$	InGMM	SVM
Airplanes	**97.78**	96.41	96.62	93.22	93.69
	(0.78)	(0.74)	(0.75)	(0.80)	(0.79)
p-value	N/A	7.71×10^{-4}	0.0031	1.48×10^{-10}	7.57×10^{-10}
Faces	**94.92**	93.37	93.62	89.42	89.60
	(0.56)	(1.97)	(0.98)	(1.70)	0.65
p-value	N/A	0.028	0.002	1.46×10^{-8}	1.37×10^{-13}
Cars	**99.26**	97.85	97.97	94.68	97.25
	(0.64)	(1.13)	(0.82)	(0.73)	(0.68)
p-value	N/A	0.0029	9.57×10^{-4}	1.28×10^{-11}	2.31×10^{-6}
Motorbikes	**94.31**	93.03	93.24	90.24	89.29
	(0.63)	(0.89)	(0.77)	(0.64)	(0.83)
p-value	N/A	0.0017	0.0033	2.63×10^{-11}	9.88×10^{-12}

example, the scale-invariant feature transform (SIFT) [57] descriptor, the local binary pattern (LBP) descriptor [58], and the Histogram of Oriented Gradient (HOG) descriptor [59]. The HOG descriptor, among others, has been one of the most popular and effective one for image categorization or detection [60], [61]. In this paper, we employ the rectangular HOG (R-HOG) descriptor [62], an variant and improved version of HOG. With the principles of R-HOG and by considering seven windows and nine histogram bins, each image is represented by a 441-dimensional positive feature vector.

3) Image Categorization: Object categorization refers to classifying a given image into a specific category, such as car, face, motorbike, and airplane. It can also be considered as an image categorization problem [63], which is an important and challenging problem in a wide range of application areas such as multimedia retrieval, pattern recognition and computer vision. Image categorization and its related applications have attracted considerable attention during the past few years [64]–[68]. The reason that image categorization has emerged as one of the most active areas in the fields of image understanding and computer vision is mainly because its large potential in web image research, video retrieval, image database annotation, and medical image mining. Although human usually perform well on the task of image categorization, it remains difficult for computers to achieve similar performance. This is due to the various poses, different scales, multiple viewpoints.

Our experiments for image categorization were implemented as follows. First, R-HOG descriptors were extracted from each image. Each image in the datasets was then represented by a 441-dimensional positive vector. Second, the vectors from one category are assumed to be generated from an InIDMM. Each category has been randomly divided into equal training and test sets. For each category, one InIDMM was trained based on the training set. Third, the proposed Bayesian InIDMM was employed as a classifier to categorize objects by assigning the test image to a given class that has the highest posterior probability. Table III lists the average categorization accuracies. It can be observed that the proposed InIDMM$_{\text{SLB}}$ is superior to all the other referred methods. In order to remove the randomness effect in the results, we conducted 10 rounds of simulations and the mean values with the standard deviations are reported. The accuracy distributions are shown in Fig. 8.

4) Object Detection: Object detection is another essential problem in computer vision and has been commonly applied in various applications like content-based image retrieval, intelligent traffic management, driver assistance system, and video surveillance [69], [70]. The main goal of object detection is to find instances of real-world objects such as car, face, or bicycle in an images or a video clip. Typical object detection algorithms apply the extracted features and employ the



(a) Airplane

(b) Face

(c) Car

(d) Motorbike

Fig. 10: Boxplots for comparisons of the detection accuracies' distributions for the Caltech-4. The central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles. The outliers are marked individually.

learning algorithms to recognize the instances from an object class. Here, we apply the proposed InIDMM as a classifier and study its performance in object detection. Similar as image categorization, we also applied the R-HOG descriptor to represent an image. Each image in the dataset was represented by a 441-dimensional positive feature vector.

For the experiments on the Caltech-4 dataset, we evaluated the detection performance on the four sub-datasets mentioned in Sec. IV-B3. In addition these four datasets, we used the Caltech background sub-dataset (451 images) as the non-object sub-dataset for these four object sub-classes. Samples images from each of these four object classes and the Caltech background dataset are shown in Fig 7.

The proposed InIDMM is utilized as a classifier to detect the objects through assigning the testing image to a given group (object or non-object). Table IV summarizes the detection accuracies. It can be observed from these results that the InIDMM provides the best detection accuracies compared to the other methods. During the evaluations, each of the aforementioned sub-datasets were randomly into two separate halves, one for training and the other one for test. Ten rounds of simulations were conducted and the mean values with the standard deviations are reported. Figure 10 illustrates the distributions of the detection accuracies.

5) Computational efficiency: As emphasized at the introduction section of this paper, one motivation of applying the EVI framework to derive analytically tractable solution for InIDMM such that the computational cost can be reduced, compared with numerical solution. In Tab. V, we compare the required runtime for InIDMM$_{\text{SLB}}$ and InIDMM$_{\text{MCMC}}$. Ten rounds of simulations were conducted and the mean values are reported. The p-values of the student's t-test with the null-hypothesis that the runtimes of InIDMM$_{\text{SLB}}$ and InIDMM$_{\text{MCMC}}$ have equal means but unknown variances are listed.It can be concluded that the proposed InIDMM$_{\text{SLB}}$ has statistically significantly superior performance in terms of runtime.

V. Conclusions

The inverted Dirichlet distribution has been widely applied in modeling the positive vector (vector that contains only positive elements). The Dirichlet processing mixture of the inverted Dirichlet mixture model (InIDMM) can provide good modeling performance to the positive vectors. Compared to the conventional finite inverted Dirichlet mixture model (IDMM), the InIDMM has more flexible model complexity as the number of mixture components can be automatically determined. Moreover, the over-fitting and under-fitting problem is avoided by the Bayesian estimation of InIDMM. To obtain an analytically tractable solution for Bayesian estimation of InIDMM, we utilized the recently proposed extended variational inference (EVI) framework. With single lower bound (SLB) approximation, the convergence of the proposed analytically tractable solution is guaranteed, while the solution obtained via multiple lower bound (MLB) approximations may result in oscillations of the objective function. Extensive synthesized data evaluations and real data evaluations demonstrated the superior performance of the proposed method.

References

[1] B. Everitt and D. Hand, *Finite Mixture Distributions*. Chapman and Hall, London, UK, 1981.

TABLE V: Comparisons of runtime (in s)† for InIDMM$_{\text{SLB}}$ and InIDMM$_{\text{MCMC}}$.

	Image categorization		Object detection			
	ETH	Caltech-4	Airplanes	Faces	Cars	Motorbikes
InIDMM$_{\text{SLB}}$	192.02	115.95	41.56	36.97	42.30	46.53
InIDMM$_{\text{MCMC}}$	342.41	118.11	66.75	54.20	59.43	66.03
p-value	1.23×10^{-7}	5.53×10^{-9}	1.99×10^{-5}	2.46×10^{-5}	4.29×10^{-4}	1.21×10^{-4}

† On a ThinkCentre$^{\circledR}$ computer with Intel$^{\circledR}$ Core$^{\text{TM}}$ i5 $-$ 4590 CPU 8G.

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

[3] 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.

[4] S. Bram, "Modeling and analysis of wireless channels via the mixture of Gaussian distribution," vol. 65, no. 3, pp. 951–957, 2015.

[5] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 876–89, 2015.

[6] 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.

[7] 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. 2160–73, 2011.

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

[9] N. Nasios and A. G. Bors, "Variational learning for Gaussian mixture models," *IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics)*, vol. 36, no. 4, pp. 849–862, July 2006.

[10] S. Sun and X. Xu, "Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 466–475, June 2011.

[11] J. Jung, S. R. Lee, H. Park, S. Lee, and I. Lee, "Capacity and error probability analysis of diversity reception schemes over generalized-K fading channels using a mixture gamma distribution," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 4721–4730, Sept 2014.

[12] J. Taghia, Z. 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, Sept 2014.

[13] E. A. Houseman, B. C. Christensen, R. F. Yeh, C. J. Marsit, M. R. Karagas, M. Wrensch, H. H. Nelson, J. Wiemels, S. Zheng, J. K. Wiencke, and K. T. Kelsey, "Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions," *Bioinformatics*, vol. 9, p. 365, 2008.

[14] J. M. P. Nascimento and J. M. Bioucas-Dias, "Hyperspectral unmixing based on mixtures of Dirichlet components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 863–878, March 2012.

[15] Q. He, K. Chang, E. P. Lim, and A. Banerjee, "Keep it simple with time: A reexamination of probabilistic topic detection models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1795–1808, Oct 2010.

[16] 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.

[17] ——, "Bayesian learning of inverted Dirichlet mixtures for SVM kernels generation," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1443–1458, 2013.

[18] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, no. 11, pp. 1457–1469, Nov. 2004.

[19] S. C. Markley and D. J. Miller, "Joint parsimonious modeling and model order selection for multivariate Gaussian mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 548–559, June 2010.

[20] Z. Liang and S. Wang, "An EM approach to MAP solution of segmenting tissue mixtures: a numerical analysis." *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 297–310, 2009.

[21] N. Bouguila and D. Ziou, "Unsupervised selection of a finite Dirichlet mixture model: an MML-based approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 993–1009, June 2006.

[22] S. Richardson and P. J. Green, "Corrigendum: On bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society*, vol. 60, no. 3, p. 661, 1996.

[23] S. Sun, "A review of deterministic approximate inference techniques for Bayesian machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2039–2050, Dec. 2013.

[24] L. Huang, Y. Xiao, K. Liu, H. C. So, and J. K. Zhang, "Bayesian information criterion for source enumeration in large-scale adaptive antenna array," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3018–3032, May 2016.

[25] X. Chen, "Using Akaike information criterion for selecting the field distribution in a reverberation chamber," *IEEE Transactions on Electromagnetic Compatibility*, vol. 55, no. 4, pp. 664–670, Aug 2013.

[26] K. Bousmalis, S. Zafeiriou, L. P. Morency, M. Pantic, and Z. Ghahramani, "Variational infinite hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1917–1929, Sept 2015.

[27] M. Meilă and H. Chen, "Bayesian non-parametric clustering of ranking data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2156–2169, Nov 2016.

[28] Y. Xu, M. Megjhani, K. Trett, W. Shain, B. Roysam, and Z. Han, "Unsupervised profiling of microglial arbor morphologies and distribution using a nonparametric Bayesian approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 1, pp. 115–129, Feb 2016.

[29] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

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

[31] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, Eds., *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[32] Y. W. Teh and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[33] N. J. Foti and S. A. Williamson, "A survey of non-exchangeable priors for Bayesian nonparametric models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 359–371, Feb 2015.

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

[35] X. Wei and C. Li, "The infinite student's t -mixture for robust modeling," *Signal Processing*, vol. 92, no. 1, pp. 224–234, 2012.

[36] N. Bouguila and D. Ziou, "A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2010.

[37] X. Wei and Z. Yang, "The infinite student's t -factor mixture analyzer for robust clustering and classification ," *Pattern Recognition*, vol. 45, no. 12, pp. 4346–4357, 2012.

[38] S. P. Chatzis and G. Tsechpenakis, "The infinite hidden Markov random field model." *IEEE Transactions on Neural Networks*, vol. 21, no. 6, pp. 1004–14, 2010.

[39] M. Wedel and P. Lenk, *Markov Chain Monte Carlo*. Boston, MA: Springer US, 2013, pp. 925–930.

[40] M. Pereyra, P. Schniter, E. Chouzenoux, J. C. Pesquet, J. Y. Tourneret, A. O. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 224–241, Mar. 2016.

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

[42] 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, 2015.

[43] M. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[44] R. Ranganath, C. Wang, D. M. Blei, and E. P. Xing, “An adaptive learning rate for stochastic variational inference,” in *Proceedings of International Conference on Machine Learning*, Feb. 2013, pp. 298–306.

[45] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, “Nested hierarchical Dirichlet processes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 256–270, Feb. 2015.

[46] W. Fan and N. Bouguila, “Topic novelty detection using infinite variational inverted Dirichlet mixture models,” in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2015, pp. 70–75.

[47] B. A. Frigyik, A. Kapila, and M. R. Gupta, “Introduction to the Dirichlet distribution and related processes,” Department of Electrical Engineering, University of Washington, Tech. Rep., 2010.

[48] J. Sethuraman, “A constructive definition of the Dirichlet prior,” *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1991.

[49] X. Wei and C. Li, “The student’s t-hidden Markov model with truncated stick-breaking priors,” *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 355–358, June 2011.

[50] J. Paisley and L. Carin, “Hidden Markov models with stick-breaking priors,” *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3905–3917, June 2009.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[52] A. Corduneanu and C. M. Bishop, “Variational bayesian model selection for mixture distribution,” in *Proceedings of the* 8^{th} *International Conference on AI and Statistics*, 2001, pp. 27–34.

[53] Y. Lai, Y. Ping, B. Wang, J. Wang, and X. Zhang, “Variational Bayesian inference for finite inverted Dirichlet mixture models and its application to object detection,” *Chinese Journal of Electronics*, 2017, accepted.

[54] T. Bdiri and N. Bouguila, “An infinite mixture of inverted Dirichlet distributions,” in *International Conference on Neural Information Processing*, 2011, pp. 71–78.

[55] T. S. F. Haines and T. Xiang, “Background subtraction with Dirichlet process mixture models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 670–683, Apr. 2014.

[56] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transaction on Intelligent System Technology*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[57] L. Seidenari, G. Serra, A. D. Bagdanov, and A. D. Bimbo, “Local pyramidal descriptors for image recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1033–1040, May 2014.

[58] X. Qi, R. Xiao, C. G. Li, Y. Qiao, J. Guo, and X. Tang, “Pairwise rotation invariant co-occurrence local binary pattern,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2199–2213, Nov 2014.

[59] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Internaional Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2005, pp. 886–893.

[60] 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, Nov 2016.

[61] X. Ma, W. A. Najjar, and A. K. Roy-Chowdhury, “Evaluation and acceleration of high-throughput fixed-point object detection on FPGAs,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 6, pp. 1051–1062, June 2015.

[62] O. L. Junior, D. Delgado, V. Goncalves, and U. Nunes, “Trainable classifier-fusion schemes: An application to pedestrian detection,” in *IEEE International Conference on Intelligent Transportation Systems*, Oct 2009, pp. 1–6.

[63] L. Zhang, R. Hong, Y. Gao, R. Ji, Q. Dai, and X. Li, “Image categorization by learning a propagated graphlet path.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 674–685, 2016.

[64] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[65] H. L. Luo, H. Wei, and L. L. Lai, “Creating efficient visual codebook ensembles for object categorization,” *IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans*, vol. 41, no. 2, pp. 238–253, 2011.

[66] L. Wu, Y. Hu, M. Li, N. Yu, and X. S. Hua, “Scale-invariant visual language modeling for object categorization,” *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 286–294, 2009.

[67] J. Stottinger, A. Hanbury, N. Sebe, and T. Gevers, “Sparse color interest points for image retrieval and object categorization,” *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2681–2692, 2012.

[68] T. Deselaers, G. Heigold, and H. Ney, “Object classification by fusing SVMs and Gaussian mixtures,” *Pattern Recognition*, vol. 43, no. 7, pp. 2476–2484, 2010.

[69] T. Xiao, C. Zhang, and H. Zha, “Learning to detect anomalies in surveillance video,” *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, Sept 2015.

[70] S. J. Krotosky and M. M. Trivedi, “Person surveillance using visual and infrared imagery,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1096–1105, Aug 2008.

Zhanyu Ma has been an Associate Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He is also an adjunct Associate Professor at Aalborg University, Aalborg, Denmark, since 2015. He received his Ph.D. degree in Electrical Engineering from KTH-Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics. He is a senior member of IEEE.

Yuping Lai has been a lecturer at North China University of Technology, China, since 2014. 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.

Yi-Zhe Song is a senior lecturer at School of Electronic Engineering and Computer Science, Queen Mary, University of London. He researches into computer vision, computer graphics and their convergence, particularly perceptual grouping, image segmentation (description),cross-domain image analysis, non-photo realistic rendering, with a recent emphasis on human sketch representation, recognition and retrieval. He received both the B.Sc.(first class) and Ph.D. degrees in Computer Science from the Department of Computer Science, University of Bath, UK, in 2003 and 2008, respectively; prior to his doctoral studies, he obtained a Diploma (M.Sc.) degree in Computer Science from the Computer Laboratory, University of Cambridge, UK, in 2004. Prior to 2011, he worked at University of Bath as a Research and Teaching Fellow. He is an Associate Editor of Neurocomputing and member of IEEE and BMVA.

Liang Wang received both the B. Eng. and M. Eng. degrees from Anhui University in 1997 and 2000 respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CAS) in 2004. Currently, he is a full Professor of Hundred Talents Program at the NLPR, Institute of Automation, CAS, China. His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly-ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV and ICDM.

Bastiaan Kleijn received the PhD degree in electrical engineering from Delft University of Technology, The Netherlands (TU Delft); the MSEE degree from Stanford University, CA; and the MSc degree in physics and the PhD degree in soil science from the University of California, Riverside. He is currently a professor at Victoria University of Wellington (VUW), New Zealand, and TU Delft, The Netherlands (part time). He was a professor and head of the Sound and Image Processing Laboratory at KTH-Royal Institute of Technology, Stockholm, Sweden, from 1996 until 2010, and a founder of Global IP Solutions, a company that provided the original audio technology to Skype and was later acquired by Google. Before 1996, he was with the Research Division of AT&T Bell Laboratories in Murray Hill, New Jersey. He is an IEEE fellow.

Jun Guo received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published over 200 papers on the journals and conferences including SCIENCE, Nature Scientific Reports, IEEE Trans. on PAMI, Pattern Recognition, AAAI, CVPR, ICCV, SIGIR, etc.

Dear Editor in Chief,

Thank you for your effort in organizing the review of our manuscript.

In our previous work, we have applied the extended variational inference (EVI) framework to several non-Gaussian statistical models and demonstrated the good performance. The EVI framework, especially when applying to non-Gaussian statistical models, shows advantages over the conventional ML estimation based methods and draws more and more attentions.

In this manuscript, based on the EVI framework, we derived an analytically tractable solution for variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet mixture model and demonstrated the advantages of the proposed method.

The key contributions of our work are three-fold:

1) The finite inverted Dirichlet mixture model (IDMM) has been extended to the infinite inverted Dirichlet mixture model (InIDMM) under the stick-breaking framework [1], [2]. Thus, the difficulty in automatically learning the number of mixture components can be overcome;

2) An analytically solution is derived with the EVI framework for InIDMM., based on single lower bound approximation. Moreover, comparing with the recently proposed algorithm for InIDMM [3], which is based on multiple lower bound approximation, our algorithm can not only theoretically guarantee convergence but also provide better approximations;

3) The proposed method has been applied in several important applications, such as image categorization and object detection. The good performance has been illustrated with both synthesized and real data evaluations.

We recommend Prof. Siliang Sun to be the AE to handle the review process of our submission, as his expertise area is in Bayesian nonparametric learning, which is related to our research such that he will be familiar and provide fair judgement to this work.

Thanks again!

Best regards,

Zhanyu Ma on behalf of all the authors

[1] Y. W. Teh and D. M. Blei, "Hierarchical Dirichlet processes," Journal of the American Statistical Association, vol. 101, no. 476, pp. 1566–1581, 2006.

[2] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical Dirichlet processes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 2, pp. 256–270, Feb 2015.

[3] W. Fan and N. Bouguila, "Topic novelty detection using infinite variational inverted Dirichlet mixture models," in IEEE International Conference on Machine Learning and Applications (ICMLA), Dec 2015, pp. 70–75.