**Variational Bayesian Learning for Dirichlet Process Mixture of Inverted Dirichlet Distributions**

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>IEEE Transactions on Neural Networks and Learning Systems</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>TNNLS-2018-P-8972</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>21-Jan-2018</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Ma, Zhanyu; Beijing University of Posts and Telecommunications, Pattern Recognition and Machine Intelligence 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 Intelligence Lab.</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Dirichlet process mixture, inverted Dirichlet distribution, Bayesian estimation, variational learning, computer vision</td>
</tr>
</tbody>
</table>
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 Dirichlet 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-$K$ ($K_C$) 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...
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 poste-
rior 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 method1. 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 nonpara-
metric [31] model selection method is based on the Dirichlet
process (DP) mixture model [32], [33]. The DP mixture model1
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 mix-
ture 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 pro-
posed 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 inference (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 algorithm for InIDMM [46], [47], which
be 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 contribu-
tions 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: Sec-
tion 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-

1Here, 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]

\[
\text{IDMM}(\vec{x} | \vec{\pi}, \Lambda) = \sum_{m=1}^{M} \pi_m \text{iDir}(\vec{x} | \vec{\alpha}_m),
\]

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, $\text{iDir}(\vec{x} | \vec{\alpha})$ is an inverted Dirichlet distribution with its $(D + 1)$-dimensional positive parameter vector $\vec{\alpha} = \{\alpha_1, \cdots, \alpha_{D+1}\}$ defined as

\[
\text{iDir}(\vec{x} | \vec{\alpha}) = \frac{\Gamma\left(\sum_{d=1}^{D+1} \alpha_d\right)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d-1}\left(1 + \sum_{d=1}^{D} x_d\right)^{-\sum_{d=1}^{D+1} \alpha_d},
\]

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

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 $\varphi$ is a positive real scalar. We consider two countably infinite collections of independently generated stochastic variables $\Omega_m \sim H$ and $\lambda_m \sim \text{Beta}(\lambda_m; 1, \varphi)$ for $m = \{1, \cdots, \infty\}$, where Beta$(\cdot; a, b)$ is the beta distribution defined as 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 $\varphi$ and a base measure or base distribution $H$ (denoted as $G \sim \text{DP}(\varphi, H)$), if the following conditions are satisfied:

\[
G = \sum_{n=1}^{\infty} \pi_n \delta_{\Omega_n}, \quad \pi_m = \lambda_m \prod_{l=1}^{m-1} (1 - \lambda_l),
\]

where $\{\pi_m\}$ is a set of stick-breaking weights with constraints $\sum_{m=1}^{\infty} \pi_m = 1$, $\delta_{\Omega_m}$ is a delta function whose value is 1 at location $\Omega_m$ 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

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.

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

\[
\text{InIDMM}(\vec{x} | \vec{\pi}, \Lambda) = \sum_{m=1}^{\infty} \pi_m \text{iDir}(\vec{x} | \vec{\alpha}_m),
\]

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 $X = \{\vec{x}_n\}_{n=1}^{N}$ is given by

\[
\text{InIDMM}(X | \vec{\pi}, \Lambda) = \prod_{n=1}^{N} \left\{ \sum_{m=1}^{\infty} \pi_m \text{iDir}(\vec{x}_n | \vec{\alpha}_m) \right\}.
\]

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, $z_{nn}$ is indicative to the sample realization $\vec{x}_n$ comes from the mixture component $m$. Therefore, the conditional distribution of $X$ given the parameters $\Lambda$ and the latent variables $Z$ is

\[
\text{InIDMM}(X | Z, \Lambda) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \text{iDir}(\vec{x}_n | \vec{\alpha}_m)^{z_{nm}}.
\]

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 $Z$, $\Lambda$, and $\varphi$ such that a full Bayesian estimation model can be obtained. In the aforementioned full Bayesian model, the prior distribution of $Z$ given $\vec{\pi}$ is given by

\[
p(Z | \vec{\pi}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \pi_m^{z_{nm}}.
\]
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
\[
p(\vec{Z} | \vec{\lambda}) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \lambda_m \prod_{t=1}^{m-1} (1 - \lambda_t)^{Z_{nt}-1}.
\] (8)

As previously mentioned in Section II-B, the prior distribution of $\vec{\lambda}$ is
\[
p(\vec{\lambda} | \vec{\varphi}) = \prod_{m=1}^{\infty} \text{Beta}(\lambda_m; 1, \varphi_m) = \prod_{m=1}^{\infty} \varphi_m (1 - \lambda_m)^{\varphi_m - 1},
\] (9)
where $\vec{\varphi} = (\varphi_1, \varphi_2, \ldots)$. Based on (3), we can obtain the expected value of $\pi_m$. In order to do this, the expected value of $\lambda_m$ will first be calculated as
\[
\langle \lambda_m \rangle = 1/(1 + \varphi_m).
\] (10)
Then, the expected value of $\pi_m$ is denoted as
\[
\langle \pi_m \rangle = \langle \lambda_m \rangle \prod_{t=1}^{m-1} (1 - \langle \lambda_t \rangle).
\] (11)
It is worth to note that, when the value of $\varphi_m$ is small, $\langle \lambda_m \rangle$ will become large. Therefore, the expected value of $\pi_m$ will yield small $\pi_m$ that the distribution of $\pi_m$ will be sparse.

As $\varphi_m$ is positive, assume $\vec{\varphi}$ follows a product of gamma prior distributions as
\[
p(\vec{\varphi}; \vec{s}, \vec{t}) = \prod_{m=1}^{\infty} \text{Gam}(\varphi_m; s_m, t_m) = \prod_{m=1}^{\infty} \frac{\Gamma(s_m)^{\psi_m}}{\Gamma(s_m)} e^{-t_m \varphi_m},
\] (12)
where $\text{Gam}(\cdot)$ is the gamma distribution, $\vec{s} = (s_1, s_2, \ldots)$ and $\vec{t} = (t_1, t_2, \ldots)$ are the hyperparameters and subject to the constraints $s_m > 0$ and $t_m > 0$.

Next, we introduce an approximating conjugate prior distribution to parameter $\Lambda$ in InDDM. The inverted Dirichlet distribution belongs to the exponential family and its formal conjugate prior can be derived with the Bayesian rule [2] as
\[
p(\vec{\alpha} | \mu_0, v_0) = C(\mu_0, v_0) \left[ \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \right] e^{-\sum_{d=1}^{D+1} \alpha_d (\vec{\alpha}^T - \vec{I}_{D+1})},
\] (13)
where $\vec{\mu}_0 = [\mu_0, \ldots, \mu_{D+1}]$ and $v_0$ are the hyperparameters in the prior distribution, $C(\mu_0, v_0)$ is a normalization coefficient such that $\int p(\vec{\alpha} | \mu_0, v_0) d\vec{\alpha} = 1$. $\vec{I}_D$ 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 $\vec{X}$)
\[
f(\vec{\alpha} | \vec{X}) = \frac{\text{iDir}(\vec{X} | \vec{\alpha}) f(\vec{\alpha})}{\int \text{iDir}(\vec{X} | \vec{\alpha}) f(\vec{\alpha}) d\vec{\alpha}}
= C(\vec{\mu}_N, \nu_N) \left[ \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \right] e^{-\sum_{d=1}^{D+1} \alpha_d (\vec{\alpha}^T - \vec{I}_{D+1})},
\] (14)
where the hyperparameters $\nu_N$ and $\vec{\mu}_N$ in the posterior distribution are
\[
\nu_N = v_0 + N, \vec{\mu}_N = \vec{\mu}_0 - \ln \vec{X}^+ - \vec{I}_{D+1} \ln (1 + \vec{I}_{D+1} \vec{X}^+) / \nu_N.
\] (15)
In (15), $\vec{X}^+$ is a $(D + 1) \times N$ matrix by connecting $\vec{I}_{D+1}$ to the bottom of $\vec{X}$. However, it is not applicable in our VI framework due to the analytically intractable normalization factor in (44). Because $\Lambda$ is positive, we adopt gamma prior distributions to approximate conjugate prior for $\Lambda$ as well. By assuming the parameters of inverted Dirichlet distribution are mutually independent, we have
\[
p(\Lambda; U, V) = \prod_{m=1}^{\infty} \prod_{d=1}^{D+1} \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} e^{-v_m \lambda_m^{u_m - 1}},
\] (16)
where all the hyperparameters $U = \{u_m\}$ and $V = \{v_m\}$ are positive.

With the Bayesian rules and by combining (6) and (8)-(16) together, we can represent the joint density of the observation $\vec{X}$ with all the i.i.d. latent variables $\Theta = (\vec{Z}, \Lambda, \vec{\lambda}, \vec{\varphi})$ as
\[
p(\vec{X}, \Theta) = p(\vec{X} | \vec{Z}, \Lambda)p(\vec{Z} | \vec{\lambda})p(\vec{\lambda} | \vec{\varphi})p(\vec{\varphi})
= \prod_{n=1}^{N} \prod_{m=1}^{\infty} \lambda_m \prod_{d=1}^{D+1} \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} e^{-v_m \lambda_m^{u_m - 1}}
\times \prod_{m=1}^{\infty} \prod_{d=1}^{D+1} \lambda_m \prod_{d=1}^{D+1} \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} e^{-v_m \lambda_m^{u_m - 1}}
\times \prod_{m=1}^{\infty} \prod_{d=1}^{D+1} \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} e^{-v_m \lambda_m^{u_m - 1}}
\] (17)
The structure of the InDDM 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 INDDM

In this section, we develop a variational Bayesian inference framework for learning the InDDM. 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

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

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 \(E(q)[\ln p(X, \Theta)]\) is analytically intractable such that the lower bound \(L(q)\) 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 \(\tilde{p}(X, \Theta)\) that satisfies

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

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

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

If \(\tilde{p}(X, \Theta)\) 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

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

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 \(p(\Theta|X)\) as

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

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 opposite 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 \(\tilde{L}(q)\) shown in (20). The optimal solution in this case is given by

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

where \(\langle \cdot \rangle_{\hat{j} \neq s}\) 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 \(\Theta_s\) 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

\[
\langle \ln p(X, \Theta) \rangle = \sum_{m=1}^{M} \sum_{d=1}^{D} \langle z_{nm} \rangle \left[ R_m + \sum_{d=1}^{D} \left( \langle \alpha_{md} \rangle - 1 \right) \ln x_{nd} \right.
\]

\[
- \sum_{d=1}^{D} \langle \alpha_{md} \rangle \left( 1 + \sum_{d=1}^{D} x_{nd} + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln (1 - \lambda_j) \rangle \right) \]

\[
+ \sum_{m=1}^{M} \left[ (\ln \varphi_m) + (\langle \varphi_m \rangle - 1)(\ln (1 - \lambda_m)) \right] \]

\[
+ \sum_{m=1}^{M} \sum_{d=1}^{D} \left[ (u_{md} - 1)(\ln \alpha_{md} - v_{md}\langle \alpha_{md} \rangle) \right]
\]

\[
+ \sum_{m=1}^{M} \left[ (s_{m} - 1)(\ln \varphi_m) - t_m \langle \varphi_m \rangle \right] + \text{Con.}, \tag{24}
\]

where \(R_m = \langle \ln \Gamma(\sum_{d=1}^{D} \langle \alpha_{md} \rangle) \rangle \). With the mathematical expression in (24), an analytically tractable solution is not feasible, which is due to the fact that \(R_m\) 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 \(\hat{R}_m\) such that \(\hat{R}_m \geq R_m\). According to [6, Eq. 25], we can select \(\hat{R}_m\) as

\[
\hat{R}_m = \ln \prod_{d=1}^{D} \frac{\Gamma(\langle \alpha_{md} \rangle)}{\Gamma(\langle \alpha_{md} \rangle) + \sum_{d=1}^{D} \left[ \Psi(\sum_{d=1}^{D} \langle \alpha_{md} \rangle) - \Psi(\langle \alpha_{md} \rangle) \right]}
\]

\[
\times \left[ (\ln \alpha_{md} - \langle \alpha_{md} \rangle) \right] + \text{Con.}, \tag{25}
\]

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 \(\langle \ln p(X, \Theta) \rangle\) can be obtained as

\[
\langle \ln \hat{p}(X, \Theta) \rangle = \sum_{m=1}^{M} \sum_{d=1}^{D} \langle z_{nm} \rangle \left[ \hat{R}_m + \sum_{d=1}^{D} \left( \langle \alpha_{md} \rangle - 1 \right) \ln x_{nd} \right.
\]

\[
- \sum_{d=1}^{D} \langle \alpha_{md} \rangle \left( 1 + \sum_{d=1}^{D} x_{nd} + \langle \ln \lambda_m \rangle + \sum_{j=1}^{m-1} \langle \ln (1 - \lambda_j) \rangle \right) \]

\[
+ \sum_{m=1}^{M} \left[ (\ln \varphi_m) + (\langle \varphi_m \rangle - 1)(\ln (1 - \lambda_m)) \right] \]

\[
+ \sum_{m=1}^{M} \sum_{d=1}^{D} \left[ (u_{md} - 1)(\ln \alpha_{md} - v_{md}\langle \alpha_{md} \rangle) \right]
\]

\[
+ \sum_{m=1}^{M} \left[ (s_{m} - 1)(\ln \varphi_m) - t_m \langle \varphi_m \rangle \right] + \text{Con.}. \tag{26}
\]

With (23), we can get analytically tractable solutions for optimally estimating the posterior distributions of \(Z, \lambda, \varphi, \) and \(\Lambda\). We now consider each of these in more detail: 1) The posterior distribution of \(q(Z)\)
As any term that is independent of $z_{nm}$ can be absorbed into the additive constant, we have
\[
\ln q^* (z_{nm}) = \text{Con.} + z_{nm} \left[ \bar{R}_m + (\ln \lambda_m) + \sum_{j=1}^{m-1} (\ln (1 - \lambda_j)) \right]
+ \sum_{d=1}^{D} ((\alpha_{md}) - 1) \ln x_{nd} + \sum_{d=1}^{D} (\alpha_{md}) \ln (1 + \sum_{d=1}^{D} x_{nd})
\]
which has same logarithmic form of the prior distribution (i.e., the categorical distribution). Therefore, we can write $\ln q^*(Z)$ as
\[
\ln q^*(Z) = \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \ln \rho_{nm} + \text{Con.}
\]
with the definition that
\[
\ln \rho_{nm} = (\ln \lambda_m) + \sum_{j=1}^{m-1} (\ln (1 - \lambda_j)) + \bar{R}_m
+ \sum_{d=1}^{D} ((\alpha_{md}) - 1) \ln x_{nd} - \sum_{d=1}^{D} (\alpha_{md}) (1 + \sum_{d=1}^{D} x_{nd}).
\]
Recalling that $z_{nm} \in (0, 1)$ and $\sum_{m=1}^{M} z_{nm} = 1$, we define
\[
r_{nm} = \frac{\rho_{nm}}{\sum_{m=1}^{M} \rho_{nm}}.
\]
Taking the exponential of both sides of (28), we have
\[
q^*(Z) = \prod_{n=1}^{N} \prod_{m=1}^{M} r_{nm}^m,
\]
which is the optimal posterior distribution of $Z$.

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 the optimal solutions to $\bar{X}$, $\bar{\varphi}$, and $\Lambda$, respectively. The derivation details can be found in the appendix.

2) The posterior distribution of $\bar{X}$

The optimal solution to the posterior distribution of $\bar{X}$ is characterized as
\[
q(\bar{X}) = \prod_{m=1}^{M} \text{Beta}(\lambda_m; g_m^*, h_m^*),
\]
where the hyperparameters $g_m^*$ and $h_m^*$ are
\[
g_m^* = 1 + \sum_{n=1}^{N} \langle z_{nm} \rangle, \quad h_m^* = \langle \varphi_m \rangle + \sum_{n=1}^{N} \sum_{j=1}^{m-1} \langle z_{nj} \rangle.
\]

3) The posterior distribution of $\bar{\varphi}$

The optimal solution to the posterior distribution of $\bar{\varphi}$ is
\[
q^*(\bar{\varphi}) = \prod_{m=1}^{M} \text{Gam}(\varphi_m; s_m^*, t_m^*),
\]
where the optimal solutions to the hyperparameters $s_m^*$ and $t_m^*$ are
\[
s_m^* = 1 + s_0^m, \quad t_m^* = t_0^m - \langle \ln (1 - \lambda_m) \rangle,
\]
where $s_0^m$ and $t_0^m$ denote the hyperparameters initialized in the prior distribution, respectively.

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

The optimal approximation to the posterior distribution of $\Lambda$ is
\[
q^*(\Lambda) = \prod_{m=1}^{M} \prod_{d=1}^{D+1} \text{Gam}(\alpha_{md}; u_{md}^*, v_{md}^*),
\]
where the optimal solutions to the hyperparameters $u_{md}^*$ and $v_{md}^*$ are given by
\[
u_{md}^* = u_{md} + \sum_{n=1}^{N} \langle z_{nm} \rangle \left[ \Psi \left( \sum_{k=1}^{K+1} (\alpha_{mk}) \right) - \Psi(\langle \alpha_{md} \rangle) \right] (\alpha_{md})
\]
and
\[
u_{md}^* = v_{md} - \sum_{n=1}^{N} \langle z_{nm} \rangle \left[ \ln x_{nd} - \ln (1 + \sum_{d=1}^{D} x_{nd}) \right].
\]

In the above equations, $u_0^d$ and $v_0^d$ 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:
\[
\langle \ln (1 - \lambda_m) \rangle = \langle \ln g_m^* + h_m^* \rangle = \langle \ln (g_m + h_m - 1) \rangle,
\]
\[
\langle \ln \alpha_{md} \rangle = \langle \Psi(u_{md}^* + v_{md}^*) \rangle - \langle \ln u_{md}^* \rangle - \langle \ln v_{md}^* \rangle.
\]

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 $\tilde{L}(q)$
\[
\tilde{L}(q) = \int q(\Theta) \ln \frac{q(\Theta; X)}{q(\Theta)} d\Theta
= \langle \ln \tilde{p}(X; \Theta) \rangle - \langle \ln q(\Theta) \rangle
= \langle \ln \tilde{p}(X; \Theta) \rangle - \langle \ln q(\bar{X}) \rangle - \langle \ln q(\bar{\varphi}) \rangle - \langle \ln q(\Lambda) \rangle,
\]
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
\[
\langle \ln q(\bar{X}) \rangle = \sum_{m=1}^{M} \ln \Gamma(g_m^* + h_m^*) - \ln \Gamma(g_m) - \ln \Gamma(h_m^*)
+ (g_m - 1) \langle \ln \lambda_m \rangle + (h_m^* - 1) \langle \ln (1 - \lambda_m) \rangle,
\]
\[
\langle \ln q(\bar{\varphi}) \rangle = \sum_{m=1}^{M} \langle s_m^* \rangle \ln t_m^* - \ln \Gamma(s_m^*)
+ (s_m^* - 1) \langle \ln \varphi_m \rangle - t_m^* \langle \varphi_m \rangle,
\]
\[
\langle \ln q(\Lambda) \rangle = \sum_{m=1}^{M} \sum_{d=1}^{D+1} \langle u_{md}^* \rangle \ln v_{md}^* - \ln \Gamma(u_m^*)
+ (u_m^* - 1) \langle \ln \alpha_{md} \rangle - v_{md}^* \langle \alpha_{md} \rangle.
\]
Algorithm 1 Algorithm for EVI-based Bayesian InIDMM

1: Set the initial truncation level $M$ and the initial values for hyperparameters $s_0^u, t_0^u, u_0^md,$ and $v_0^md$
2: Initialize the values of $r_{nm}$ by K-means algorithm.
3: repeat
4:   Calculate the expectations in (39).
5:   Update the posterior distributions for each variable by (33), (35), (37) and (38).
6: until Stop criterion is reached.
7: For all $m$, calculate $\langle \lambda_m \rangle = s_m^u/(s_m^u + t_m^u)$ and substitute it back into (11) to get the estimated values of the mixing coefficients $\tilde{\pi}_m$.
8: Determine the optimum number of components $M$ by eliminating the components with mixing weights smaller than $10^{-5}$.
9: Renormalize $\{\tilde{\pi}_m\}$ to have a unit $l_1$ norm.
10: Calculate $\hat{\alpha}_{md} = u_{md}^*/v_{md}^*$ for all $m$ and $d$.

Additionally, $\langle \ln \tilde{p}(X, \Theta) \rangle$ 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_{SLB}) with the Bayesian InIDMM using the MLB approximation (proposed in [46] and denoted as InIDMM_{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_{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 variational lower-bounds and the number of mixture components in Fig. 4.

2) Observations of Oscillations: We ran the InIDMM_{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_{MLB}. Figure 3 illustrates the decreasing values during iterations. On the other hand, the variational objective function obtained with the InIDMM_{SLB} algorithm was always increasing until convergence, as the SLB approximation insures the convergence 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_{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_{SLB} increases during iterations and in most cases it increases very fast.

3) Quantitative Comparisons: Next, we compare the InIDMM_{SLB} with the InIDMM_{MLB} quantitatively. With a known IDMM, 2000 samples were generated. The InIDMM_{SLB} and the InIDMM_{MLB} were applied to estimate the posterior distributions of the model, respectively. In Tab. I,
TABLE I: Comparisons of true and estimated models.

<table>
<thead>
<tr>
<th>True Model</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1 = 0.5$, $\tilde{\alpha}_1 = [16.86 2]^T$</td>
<td>$\pi_1 = 0.25$, $\tilde{\alpha}_1 = [12.36 14.18 55.16]^T$</td>
</tr>
<tr>
<td></td>
<td>$\pi_2 = 0.5$, $\tilde{\alpha}_2 = [8.12 15.18]^T$</td>
<td>$\pi_2 = 0.25$, $\tilde{\alpha}_2 = [32.48 25.12 36.48]^T$</td>
</tr>
<tr>
<td>InIDMM $\pi_1 = 0.50 \times \tilde{\alpha}_1 = [16.98 8.58 6.39 12.49]^T$</td>
<td>$\pi_1 = 0.25 \times \tilde{\alpha}_1 = [12.26 36.39 14.30 19.19 56.36 16.25]^T$</td>
<td></td>
</tr>
<tr>
<td>InIDMM $\pi_2 = 0.49 \times \tilde{\alpha}_2 = [8.20 12.16 15.49 18.34]^T$</td>
<td>$\pi_2 = 0.25 \times \tilde{\alpha}_2 = [33.37 49.92 25.85 12.80 37.00 49.79]^T$</td>
<td></td>
</tr>
<tr>
<td>MLB $\pi_1 = 0.50 \times \tilde{\alpha}_1 = [15.20 7.17 5.90 11.64]^T$</td>
<td>$\pi_1 = 0.249 \times \tilde{\alpha}_1 = [12.18 37.82 14.56 18.85 57.32 16.44]^T$</td>
<td></td>
</tr>
<tr>
<td>MLB $\pi_2 = 0.492 \times \tilde{\alpha}_2 = [9.21 13.76 17.13 21.10]^T$</td>
<td>$\pi_2 = 0.249 \times \tilde{\alpha}_2 = [33.73 51.10 26.92 12.89 38.66 51.79]^T$</td>
<td></td>
</tr>
<tr>
<td>MLB $\pi_3 = 0.25 \times \tilde{\alpha}_3 = [8.14 28.94 16.72 33.46 12.32 25.20]^T$</td>
<td>$\pi_3 = 0.25 \times \tilde{\alpha}_3 = [5.82 27.43 15.77 31.14 11.82 23.58]^T$</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model &amp; Method</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>InIDMM $\pi_1 = 0.20 \times \tilde{\alpha}_1 = [12.08 20.89 36.25 18.28 32.69 65.72 76.70]^T$</td>
<td>$\pi_2 = 0.2 \times \tilde{\alpha}_2 = [28.42 21.85 41.48]^T$</td>
<td>$\pi_3 = 0.2 \times \tilde{\alpha}_3 = [53.12 18.44 65.35 52]^T$</td>
</tr>
<tr>
<td></td>
<td>InIDMM $\pi_1 = 0.20 \times \tilde{\alpha}_1 = [5.82 27.43 15.77 31.14 11.82 23.58]^T$</td>
<td>$\pi_2 = 0.2 \times \tilde{\alpha}_2 = [28.42 21.85 41.48]^T$</td>
<td>$\pi_3 = 0.2 \times \tilde{\alpha}_3 = [53.12 18.44 65.35 52]^T$</td>
</tr>
<tr>
<td></td>
<td>InIDMM $\pi_1 = 0.20 \times \tilde{\alpha}_1 = [12.56 21.50 37.69 19.00 33.06 68.04 79.64]^T$</td>
<td>$\pi_2 = 0.20 \times \tilde{\alpha}_2 = [28.26 43.02 20.85 8.14 55.36 21.21 49.17]^T$</td>
<td>$\pi_3 = 0.20 \times \tilde{\alpha}_3 = [63.61 45.48 32.02 66.63 74.31 15.21 45.45]^T$</td>
</tr>
<tr>
<td></td>
<td>InIDMM $\pi_1 = 0.20 \times \tilde{\alpha}_1 = [12.56 21.50 37.69 19.00 33.06 68.04 79.64]^T$</td>
<td>$\pi_2 = 0.20 \times \tilde{\alpha}_2 = [28.26 43.02 20.85 8.14 55.36 21.21 49.17]^T$</td>
<td>$\pi_3 = 0.20 \times \tilde{\alpha}_3 = [63.61 45.48 32.02 66.63 74.31 15.21 45.45]^T$</td>
</tr>
</tbody>
</table>

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

we list the estimated parameters by taking the posterior means. It can be observed that, both the InIDMM $\text{SLB}$ and the InIDMM $\text{MLB}$ 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 $\text{SLB}$ over the InIDMM $\text{MLB}$ can be observed from Tab. II. The mean values of the objective function obtained by InIDMM $\text{SLB}$ are larger than those obtained by the InIDMM $\text{MLB}$ while the computational cost (measured in seconds) required by the InIDMM $\text{SLB}$ are smaller than those required by the InIDMM $\text{MLB}$. Moreover, smaller KL divergences of the estimated models from the corresponding true models also verify that the InIDMM $\text{SLB}$ yields better estimates than the InIDMM $\text{MLB}$. In order to examine if the differences between the InIDMM $\text{SLB}$ and the InIDMM $\text{MLB}$ are statistically significant, we conducted the student’s t-test with the null-

$^4$ Here, the KL divergence is calculated as $\text{KL}(\pi(x|\Theta)||\pi(x|\tilde{\Theta}))$ by sampling method. $\tilde{\Theta}$ denotes the point estimate of the parameters from the posterior distribution.
brackets. The obtained with different models. The standard deviations are in the SLB that InIDMM unknown variances are listed.

toolbox [56]). classifier (discriminant method, implemented with LIBSVM mixture model (InGMM, another commonly used statistical object detection. The referred methods for comparisons are the has been applied for the task of image categorization and

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 InIDMMSLB and the referring method have equal means but unknown variances are listed.

<table>
<thead>
<tr>
<th></th>
<th>InIDMMSLB</th>
<th>InIDMMLB</th>
<th>InIDMMSCMC</th>
<th>InGMM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech-4</td>
<td>93.49</td>
<td>89.27</td>
<td>90.21</td>
<td>83.92</td>
<td>92.72</td>
</tr>
<tr>
<td>p-value</td>
<td>(1.05)</td>
<td>(0.84)</td>
<td>(0.73)</td>
<td>(0.72)</td>
<td>(0.82)</td>
</tr>
<tr>
<td>ETH-80</td>
<td>75.49</td>
<td>72.88</td>
<td>73.05</td>
<td>68.88</td>
<td>72.47</td>
</tr>
<tr>
<td>p-value</td>
<td>(0.75)</td>
<td>(1.40)</td>
<td>(0.78)</td>
<td>(0.74)</td>
<td>(0.70)</td>
</tr>
</tbody>
</table>

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 InIDMMSLB over the InIDMMMLB 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 InIDMMSLB has been applied for the task of image categorization and object detection. The referred methods for comparisons are the IDMMLB [53], the Markov Chain Monte Carlo-based numerical model estimation (InIDMMSCMC, 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]).

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
The reason that image categorization has emerged as one attractive problem during the past few years [64]– [67]. Image categorization and its related applications have such as multimedia retrieval, pattern recognition and computer vision. Image categorization is an important problem in a wide range of application areas such as content-based image retrieval, pattern recognition and computer vision. 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 as a 441-dimensional positive feature 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 employed as a classifier to categorize objects by a posterior probability. Table III lists the average categorization accuracies. It can be observed that the proposed InIDMM was superior to all the other referred methods. In order to 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.

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 InIDMMSLB and the referring method have equal means but unknown variances are listed.

<table>
<thead>
<tr>
<th>Category</th>
<th>InIDMMSLB</th>
<th>InIDMM</th>
<th>InGMM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplanes</td>
<td>97.78</td>
<td>96.62</td>
<td>93.22</td>
<td>93.69</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0031</td>
<td>10^{-10}</td>
<td>8.57×10^{-10}</td>
<td>8.57×10^{-10}</td>
</tr>
<tr>
<td>Faces</td>
<td>94.92</td>
<td>94.47</td>
<td>93.62</td>
<td>89.42</td>
</tr>
<tr>
<td>p-value</td>
<td>0.028</td>
<td>0.002</td>
<td>1.46×10^{-6}</td>
<td>1.37×10^{-13}</td>
</tr>
<tr>
<td>Cars</td>
<td>99.20</td>
<td>97.85</td>
<td>97.70</td>
<td>94.68</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0029</td>
<td>9.57×10^{-5}</td>
<td>1.28×10^{-11}</td>
<td>2.34×10^{-6}</td>
</tr>
<tr>
<td>Motorbikes</td>
<td>94.31</td>
<td>94.93</td>
<td>97.24</td>
<td>90.24</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0017</td>
<td>0.0033</td>
<td>2.63×10^{-11}</td>
<td>19.88×10^{-12}</td>
</tr>
</tbody>
</table>

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]– [67]. 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.

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
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 InIDMMSLB and InIDMMMCMC. 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 InIDMMSLB and InIDMMMCMC have equal means but unknown variances are listed. It can be concluded that the proposed InIDMMSLB 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

<table>
<thead>
<tr>
<th>Image categorization</th>
<th>ETH</th>
<th>Caltech-1</th>
<th>Airplanes</th>
<th>Object detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>InDMMs_L</td>
<td>0.80</td>
<td>0.95</td>
<td>0.86</td>
<td>0.90</td>
</tr>
<tr>
<td>InDMMs_M</td>
<td>0.84</td>
<td>0.98</td>
<td>0.90</td>
<td>0.93</td>
</tr>
</tbody>
</table>

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

