# Insights into Multiple/Single Lower Bound Approximation for Extended Variational Inference in Non-Gaussian Structured Data Modeling

Zhanyu Ma, *Senior Member, IEEE,* Jiyang Xie, *Student Member, IEEE,* Yuping Lai, *Member, IEEE,* Jalil Taghia, *Member, IEEE,* Jing-Hao Xue, and Jun Guo

***Abstract*—For most of non-Gaussian statistical models, the data being modeled represent strongly structured properties, such as scalar data with bounded support (*e.g.*, beta distribution), vector data with unit length (*e.g.*, Dirichlet distribution), and vector data with positive elements (*e.g.*, generalized inverted Dirichlet distribution). In practical implementations of non-Gaussian statistical models, it is infeasible to find an analytically tractable solution to estimating the posterior distributions of the parameters. Variational inference (VI) is a widely used framework in Bayesian estimation. Recently, an improved framework, namely the extended variational inference (EVI), has been introduced and applied successfully to a number of non-Gaussian statistical models. EVI derives analytically tractable solutions, by introducing lower-bound approximations to the variational objective function. In this paper, we compare two approximation strategies, namely the multiple lower-bounds (MLB) approximation and the single lower-bound (SLB) approximation, which can be applied to carry out the EVI. For implementation, two different conditions, the weak and the strong conditions, are discussed. Convergence of the EVI depends on the selection of the lower-bound, regardless of the choice of weak or strong condition. We also discuss the convergence properties to clarify the differences between MLB and SLB. Extensive comparisons are made based on some EVI-based non-Gaussian statistical models. Theoretical analysis is conducted to demonstrate the differences between the weak and strong conditions. Experimental results based on real data show advantages of the SLB approximation over the MLB approximation.**

***Index Terms*—Structured data, Beyesian estimation, non-Gaussian statistical models, extended variational inference, lower-bound approximation**

## I. Introduction

Gaussian distribution is a ubiquitous probability distribution used in statistics, signal processing, and pattern recognition [1]. However, in reality data may be neither Gaussian nor safely assumed to be Gaussian [2]. In many real-life applications, the data are well-structured and, therefore, not Gaussian distributed [3]. For example, the image pixel values [4], the reviewer's rating of an item in a recommendation system [5], [6], and the DNA methylation level data [7] are distributed in a range with bounded support. The diversity gain over the $K_G$ fading [8] and the periodogram coefficients in speech enhancement [9] are semi-bounded (nonnegative). The spatial fading correlation [10] and the yeast gene expressions [11] have directional characteristics for which data are assumed to be distributed on a unit hypersphere, *i.e.*, satisfying $l_2$ unit norm. In signal processing, the acoustic noise with colored spectra [12] and the measurement noise in the state-space model [13] are heavy-tailed. In the stock market, the asymptotic behavior of the first-order autoregressive (AR) process is clearly non-Gaussian [14] and the underlying Bayesian copula model for the stock index series are similarly non-Gaussian [15]. Although the above mentioned data represent diverse characteristics, a common property is that these data *not only* have specific support ranges, *but also* have "non-bell" distribution shapes. The natural properties of a Gaussian distribution (the definition domain is unbounded and the distribution shape is symmetric) do not fit such data well. It has been found in recent studies that explicitly utilizing the non-Gaussian characteristics can significantly improve the practical performance on non-Gaussian structured data [2], [4], [7]–[9], [11]–[13], [16]. Hence, it is of particular importance and interest to make thorough studies of non-Gaussian data and non-Gaussian statistical models.

Bayesian analysis plays an essential role in parameter estimation of statistical models [17]–[22]. Unlike the conventionally used maximum-likelihood (ML) estimation [23], Bayesian estimation assumes that the parameters are random variables with prior distributions, and derives the posterior distributions of the parameters by applying the Bayes theorem [24] through combining the prior distributions with the likelihood function obtained from the observed data [17], [25]. Estimation of the posterior distribution via Bayesian estimation has several advantages over the ML estimation. Firstly, it gives probabilistic description to the parameters, rather than simple point estimates yielded by the ML estimation. This makes Bayesian estimation more robust and reliable, by including the resulting uncertainty into the estimation [18]. Secondly, it can potentially prevent the overfitting problem, which is a main drawback of the ML estimation. This robustness against overfitting comes from marginalization, by integrating out uncertainties. Last but not the least, Bayesian estimation embodies *Occam's razor* [26], which allows a model to automatically regulate the model complexity. In the ML estimation, determination of the model complexity often requires cross validations, which can

Z. Ma, J. Xie, 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.

J. Taghia is with the Department of Information Technology, Division of Systems and Control, Uppsala University, Uppsala, Sweden.

J.-H. Xue is with the Department of Statistical Science, University College London, London, United Kingdom.

The corresponding author is Z. Ma. Email: mazhanyu@bupt.edu.cn

be non-optimal and computationally costly [17].

Varitional inference (VI), among others, is a widely used strategy to infer the posterior distributions of the parameters in Bayesian analysis [17], [27]. In a fully Bayesian model where all variables are assigned with prior distributions, the task is to minimize the Kullback-Leibler (KL) divergence from the true posterior to the approximating posterior [17, Ch. 10]. In variational inference, the difficulty in minimizing the KL divergence of the true one from the approximating posteriors is cast alternatively as a less challenging task of maximizing a lower-bound defined on the marginal likelihood (model evidence). During optimization, the posterior distributions over all the variables are updated by iteratively updating one variable (or one group of variables) in turn, while keeping the other variables unchanged. VI has been successfully applied to many Gaussian models [1], [17]. However, for many non-Gaussian statistical models, maximizing the lower-bound still involves intractable moment computations, and consequently the resulting posteriors are not available in a closed-form solution. Examples of such models previously studied in the literature are: beta mixture model (BMM) [4], Dirichlet mixture model (DMM) [28], [29], generalized Dirichlet mixture model (GDMM) [30], inverted-Dirichlet Mixture Model (iDMM) [31], [32], generalized inverted-Dirichlet mixture model (GiDMM) [33], von-Mises Fisher mixture model (VMM) [11], Watson mixture model (WMM) [34], and beta-Gamma non-negative matrix factorization (BG-NMF) [5]. Numerical methods, *e.g.*, Gibbs sampling and Markov chain Monte Carlo, are usually employed to sample from the posterior distribution [27]. Numerical methods are generally computationally costly, and diagnosing their convergence can be difficult, in particular for data from a high-dimensional space [35].

Recently, an improved framework, namely the extended variational inference (EVI) [4], [5], [11], [28], [29], [36]–[38], has become popular in solving the above mentioned problem. Similar to the classic VI framework, EVI seeks an optimal approximation to the posterior distribution. The difference is that EVI relaxes the objective function (the evidence lower-bound to the marginal likelihood) by constructing a lower-bound approximation to the objective function. Maximization of the EVI lower-bound, which uses the convexity or relative convexity [39] of the objective function, can yield analytically tractable solution so that the parameter estimation is facilitated.

Although extra systematic bias has been introduced due to the lower-bound approximation, several works have demonstrated the advantages of EVI in Bayesian estimation of statistical models [4], [5], [11], [29], [37]. In Bayesian estimation of BMM, Ma et al. [4] derived an analytically tractable solution which outperforms the numerical Gibbs sampling based method. As an extension of [4], Bayesian estimation of DMM via EVI has been proposed in [28] and [29], respectively. For directional data, von-Mises Fisher distribution is an important model in several applications. Analytically tractable solution to Bayesian estimation of VMM has been proposed by using EVI [11]. In Bayesian estimation of VMM, EVI was also applied in deriving analytically tractable solution [34]. For non-negative matrix factorization (NMF), EVI was also applied in deriving analytically tractable solutions for Poisson process (discrete) NMF [40], Gamma process NMF in music recording [37], and beta-Gamma NMF for bounded support data [5].

Convergence is an important issue in parameter estimation. For VI-based methods, the objective function maximized during each iteration is convex or relatively convex [39] in terms of the target variable's posterior distribution [17]. Hence, the convergence is theoretically guaranteed. In EVI, the introduced lower-bound approximation to the objective function can be obtained via either a single extension over the whole variable group or multiple extensions, one for a subset of the whole variable group. Based on this, two lower-bound approximation strategies are obtained: one is the single lower-bound (SLB) approximation [5], [11], [29], [37] and the other is the multiple lower-bounds (MLB) approximation [4], [28]. For EVI with the SLB approximation, convergence is also guaranteed because the VI objective function is replaced by a single lower-bound called the EVI lower-bound (*i.e.*, the single lower-bound to the original objective function in VI). The EVI lower-bound is convex or relatively convex and tight to the VI objective function, and thus theoretically guaranteed to be maximized with each iteration. However, when applying EVI with the MLB approximation, the variable group is divided into different disjoint subsets and there exist different lower-bound approximations to the objective function. During each iteration, different lower-bounds, one for each variable subset, are maximized iteratively. Since the new objective function is not unique, convergence cannot be theoretically guaranteed.

In order to clarify the convergence property of the EVI framework, we will discuss and summarize the conditions required in the EVI implementation. The SLB and MLB approximations will also be analyzed and compared qualitatively and quantitatively. It is worth to note that all the models we select and study in this paper (*i.e.*, BMM, DMM, and GiDMM) are typical non-Gaussian statistical models which have *both* SLB and MLB derivations. Other non-Gaussian models, which do not have both representations, are not discussed in this paper.

## II. Variational Inference and Extended Variational Inference

### A. Variational Inference

In Bayesian estimation, a universal solution to the variational inference (VI) framework [41] is to approximate the posterior distribution by a product of several factor distributions and then update each factor distribution individually [17]. This method is the so-called factorized approximation (FA) which was developed from the mean field theory in physics [42]. With the FA method, the variational objective function that we want to maximize can be represented as

$$\begin{aligned}\mathcal{L} &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &= \mathbf{E}_{\mathbf{Z}} \left[\ln p(\mathbf{X},\mathbf{Z}) - \ln q(\mathbf{Z})\right],\end{aligned} \tag{1}$$

where $\mathbf{X}$ is the observed data, and $\mathbf{Z}$ denotes all the latent random variable and parameters. If $\mathbf{Z}$ can be (approximately) fac-

TABLE I
REQUIRED CONDITIONS FOR EVI.

| | Auxiliary function | Form of the Auxiliary Function | Systematic Gap |
|---|---|---|---|
| Strong condition | $p(\mathbf{X},\mathbf{Z}) \geq \widetilde{p}_s(\mathbf{X},\mathbf{Z})$ | $\mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] \approxeq \ln p_i(\mathbf{Z}_i)$ † | $\mathcal{G}_\mathrm{s} > \mathcal{G}_\mathrm{w}$ |
| Weak condition | $\mathbf{E}_\mathbf{Z}\left[\ln p(\mathbf{X},\mathbf{Z})\right] \geq \mathbf{E}_\mathbf{Z}\left[\ln \widetilde{p}_w(\mathbf{X},\mathbf{Z})\right]$ | | |

† "$\approxeq$" denotes that the two formulations at the LHS and RHS have the same mathematical form, up to a constant difference.

torized into $M$ disjoint groups as $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_M\}$ and we approximate the true posterior distribution $p(\mathbf{Z}|\mathbf{X})$ as

$$p(\mathbf{Z}|\mathbf{X}) \approx q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i), \tag{2}$$

the optimal solution can be written as

$$\ln q_i^*(\mathbf{Z}_i) = \mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\ln p(\mathbf{X},\mathbf{Z})\right] + \text{const.} \tag{3}$$

The operator $\mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}$ means expectation with respect to all the variables in $\mathbf{Z}$ except for $\mathbf{Z}_i$. If the optimal solution to the posterior distribution of $\mathbf{Z}_i$, which is $\ln q_i^*(\mathbf{Z}_i)$ in (3), has the same logarithmical form as the prior distribution, the conjugate match between the prior and posterior distributions are satisfied. Then we have obtained an analytically tractable solution. However, this conjugate match is not satisfied in most of the practical problems [4], [5], [28], [29]. This is due to the fact that the optimal solution depends on the expectation computed with respect to the factor distribution [17].

### B. Extended Variational Inference

In order to satisfy the conjugate match requirement, some approximations can be applied to get a nearly optimally analytically tractable solution. Braun et al. [38] considered the zeroth-order and first-order delta method for computation of moments to derive an alternative for the objective function to simplify the calculation. Blei et al. [36] proposed a correlated topic model (CTM) and used a first-order Taylor expansion to preserve a bound such that an intractable expectation was avoided. Similar ideas were also applied in [4], [11], [28], [29] for approximating the posterior distributions in BMM, DMM, and VMM, respectively. Using Jensen's inequality has become commonplace in variational inference. In [37], the concavity of the function $-x^{-1}$ and the convexity of $-\log x$ were studied and the Jensen's inequality and the first-order Taylor expansion were applied to approximate the posterior distribution. Moreover, the EVI strategy was also applied in the low rank matrix approximation area [5], where the Taylor expansion and Jensen's inequality were both applied for the purpose of deriving analytically tractable solutions.

All the aforementioned works utilized the following property. Given an auxiliary function $\widetilde{p}(\mathbf{X},\mathbf{Z})$ which satisfies

$$\mathbf{E}_\mathbf{Z}\left[\ln p(\mathbf{X},\mathbf{Z})\right] \geq \mathbf{E}_\mathbf{Z}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] \text{ for all } \mathbf{X}, \tag{4}$$

the variational objective function (see [17], pp. 465 for more details) can be lower-bounded as

$$\begin{aligned}\mathcal{L} =& \mathbf{E}_\mathbf{Z}\left[\ln p(\mathbf{X},\mathbf{Z})\right] - \mathbf{E}_\mathbf{Z}\left[\ln q(\mathbf{Z})\right] \\ \geq& \mathbf{E}_\mathbf{Z}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] - \mathbf{E}_\mathbf{Z}\left[\ln q(\mathbf{Z})\right] \\ \triangleq& \widetilde{\mathcal{L}}.\end{aligned} \tag{5}$$

Then we can maximize the EVI lower-bound $\widetilde{\mathcal{L}}$, which is a lower-bound to the original objective function $\mathcal{L}$. Since $\widetilde{\mathcal{L}}$ is tight to $\mathcal{L}$ at least at one-point, maximizing $\widetilde{\mathcal{L}}$ would similarly maximizes $\mathcal{L}$ [4], [28], [29], [37]. The approximated optimal solution in this case is written as

$$\ln \widetilde{q}_i^*(\mathbf{Z}_i) = \mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] + \text{const.} \tag{6}$$

This method is the so-called EVI framework [4], [5], [11], [28], [29], [36]–[38]. Although it introduces systematic gap when involving the lower-bound approximation, the EVI allows more flexibility when calculating intractable integrations in non-Gaussian statistical models and provides a convenient way to obtain an analytically tractable solution.

In addition, the proposed EVI framework, similar to the conventional VI framework, can also be applied to implicit distributions [43]. The implicit distributions are a family of probability models that the PDF is intractable, but there exists a way to sample from them and/or approximate expectations under them and calculate the gradients *w.r.t.* the model parameters. In principle, the EVI framework can provide a more flexible approximation to the original lower bound for the purpose of providing the tools to handle the inference of the implicit distributions.

## III. CONVERGENCE OF EVI

We first compare the weak and strong conditions quantitatively. Secondly, we intensively compare the performance of the MLB approximation-based methods with the SLB approximation-based methods. The selected models are typical and widely applied non-Gaussian statistical models.

### A. Typical non-Gaussian Statistical Models

- *Beta mixture model (BMM)*
  Following the same notation in [4], we denote a BMM with observation data $\mathbf{x} = \{x_1, \ldots, x_N\}$ and $I$ mixture components as

$$f(\mathbf{x};\boldsymbol{\pi},\mathbf{u},\mathbf{v}) = \prod_{n=1}^{N}\sum_{i=1}^{I}\pi_i \mathrm{Beta}(x_n;u_i,v_i), \tag{7}$$

  where $\pi_i$ is the mixture weight for the $i$th mixture component with $\pi_i > 0$, $\sum_{i=1}^{I}\pi_i = 1$, $\mathbf{u} = [u_1, \cdots, u_I]$, and $\mathbf{v} = [v_1, \cdots, v_I]$. $\mathrm{Beta}(x;u,v)$ is the beta distribution, which can be presented as

$$\mathrm{Beta}(x;u,v) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}x^{u-1}(1-x)^{v-1}, \ u,v>0, \tag{8}$$

  where $x \in [0,1]$ and $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. The beta mixture

model has been widely applied in modeling non-Gaussian distributed data with bounded support, such as image pixels, the ratings assigned to an item in collaborative filtering [6], and the epigenetic mark values in epigenome-wide-association studies [5], [44].

- *Dirichlet mixture model (DMM)*
  If a $K$ dimensional vector $\mathbf{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$ contains only positive values and the summation of all the $K$ elements is smaller than one, the underlying distribution of $\mathbf{x}$ could be modeled by a Dirichlet distribution. The probability density function (PDF) of a Dirichlet distribution is [29]

$$\mathrm{Dir}\left(\mathbf{x};\mathbf{u}\right) = \frac{\Gamma\left(\sum_{k=1}^{K+1} u_k\right)}{\prod_{k=1}^{K+1}\Gamma\left(u_k\right)} \prod_{k=1}^{K+1} x_k^{u_k-1}, \tag{9}$$

  where $0 < x_k < 1$, $x_{K+1} = 1 - \sum_{k=1}^{K} x_k$, $u_k > 0$ and $\mathbf{u} = [u_1, \ldots, u_{K+1}]^{\mathrm{T}}$ is the parameter vector. The shape of the Dirichlet distribution depends on the parameters. With $I$ mixture components, a DMM can be represented, given a set of $N$ *i.i.d.* observations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as

$$f\left(\mathbf{X};\mathbf{\Pi},\mathbf{U}\right) = \prod_{n=1}^{N}\sum_{i=1}^{I} \pi_i \mathrm{Dir}\left(\mathbf{x}_n;\mathbf{u}_i\right), \tag{10}$$

  where $\pi_i > 0$, $\sum_{i=1}^{I}\pi_i = 1$, $\mathbf{\Pi} = [\pi_1, \ldots, \pi_I]^{\mathrm{T}}$ is the mixture weights and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_I]$ is the parameter matrix. For modeling the data representing proportions, *e.g.*, the weighting factors in a mixture model [17], and the topic model in document analysis [45], [46], the Dirichlet distribution and the related DMM have been extensively used.
- *Generalized inverted-Dirichlet model (GiDMM)*
  Assume $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ is a set of $K$-dimensional observations, where each vector $\mathbf{x}_n$ is generated from a GiDMM with $I$ mixture components [33], [47]

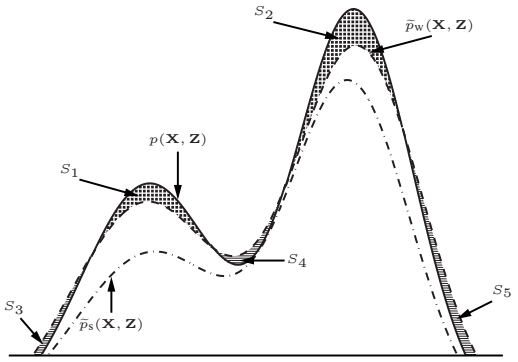
$$f\left(\mathbf{X};\mathbf{\Pi},\mathbf{U},\mathbf{V}\right) = \prod_{n=1}^{N}\sum_{i=1}^{I} \pi_i \mathrm{GiDir}(\mathbf{x}_n;\mathbf{u}_i,\mathbf{v}_i), \tag{11}$$

  where $\mathbf{\Pi} = [\pi_1, \ldots, \pi_I]^{\mathrm{T}}$ is the mixture weight vector subject to the constraints $\pi_i > 0$ and $\sum_{i=1}^{I}\pi_i = 1$, $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_I]$ and $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_I]$ are the set of parameters. $\mathrm{GiDir}(\mathbf{x};\mathbf{u},\mathbf{v})$ is a generalized inverted-Dirichlet distribution with its own positive parameter vectors $\mathbf{u} = [u_1, \cdots, u_K]^{\mathrm{T}}$ and $\mathbf{v} = [v_1, \cdots, v_K]^{\mathrm{T}}$ as

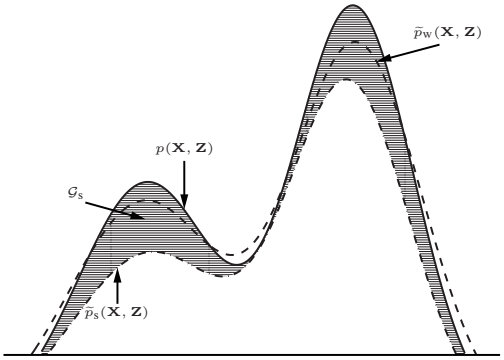
$$\mathrm{GiDir}(\mathbf{x};\mathbf{u},\mathbf{v}) = \prod_{k=1}^{K}\frac{\Gamma(u_k+v_k)}{\Gamma(u_k)\Gamma(v_k)} \cdot \frac{x_k^{u_k-1}}{\left(1+\sum_{k=1}^{K}x_k\right)^{\gamma_k}}, \tag{12}$$

  where $\Gamma(\cdot)$ represents the Gamma function, $\gamma_k = u_k + v_k - v_{k+1}$ for $k = 1, \cdots, K$ and $v_{K+1} = 0$.

For positive data which are naturally generated by several real life applications, the GiDMM, among others, has been utilized for the purpose of clustering and classification of such data [33], [48].




(a) Weak condition of EVI.




(b) Strong condition of EVI.

Fig. 1. Comparisons of the weak and strong conditions of EVI for a multi-modal distribution. The systematic gap introduced by the weak condition can be calculated as $\mathcal{G}_{\mathrm{w}} = (S_1+S_2)-(S_3+S_4+S_5)$. For either the strong or weak condition, the auxiliary function is chosen to minimize the gap as much as possible. Generally speaking, the systematic gap $\mathcal{G}_{\mathrm{w}}$ is smaller than $\mathcal{G}_{\mathrm{s}}$.

### *B. Weak Condition and Strong Condition*

As mentioned in Sec. II-B, finding an auxiliary function $\widetilde{p}(\mathbf{X},\mathbf{Z})$ is an essential yet difficult part in EVI implementation. Generally speaking, this auxiliary function should satisfy the relation presented in (4) or it should satisfy

$$p(\mathbf{X},\mathbf{Z}) \geq \widetilde{p}(\mathbf{X},\mathbf{Z}), \text{ for all } \mathbf{X} \text{ and } \mathbf{Z}. \tag{13}$$

It is straightforward to show that an auxiliary function which satisfies (13) should also satisfy (4). Hence, the condition in (4) is named as the *weak condition* and the one in (13) is referred to as the *strong condition*. When using an auxiliary function to lower-bound the original objective function, the EVI will introduce a systematic gap. Generally speaking, the gap[1] incurred by applying the weak condition is relatively smaller than that introduced by using the strong condition. Figure 1 illustrates the different gaps introduced by the weak and strong conditions, respectively.

It is worthwhile to note that the auxiliary function $\widetilde{p}(\mathbf{X},\mathbf{Z})$ is not necessary to be a normalized probability density function (PDF)[2]. This will not affect the final solution since either VI or EVI will re-normalize the obtained optimal posterior

[1]We calculate the gap via sampling methods.

[2]Actually, an auxiliary function that satisfies the strong condition cannot be a normalized PDF, as $p(\mathbf{X},\mathbf{Z})$ itself is a normalized PDF.

distribution in the end. For example, the optimal solution to $\widetilde{q}_i^*(\mathbf{Z}_i)$ can be obtained by normalizing the RHS of (6) as

$$\begin{aligned}\widetilde{q}_i^*(\mathbf{Z}_i) &= \frac{\exp\left(\ln \widetilde{q}_i^*(\mathbf{Z}_i)\right)}{\int \exp\left(\ln \widetilde{q}_i^*(\mathbf{Z}_i)\right) d\mathbf{Z}_i} \\ &= \frac{\exp\left(\mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right]\right)}{\int \exp\left(\mathbf{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right]\right) d\mathbf{Z}_i},\end{aligned} \tag{14}$$

where the constant part that does not contain $\mathbf{Z}_i$ is cancelled at the numerator and the denominator. More examples can be referred to [17, Chap. 10].

In practice, in addition to the above mentioned weak or strong condition, an auxiliary function should also have a specific mathematical form so that the optimal solution in (6) has the same logarithmic form as the prior distribution such that the conjugate match between the prior and the posterior is satisfied. This is another required condition for choosing the auxiliary function. Table I lists the required conditions when implementing EVI. In summary, in order to apply the EVI to derive an analytically tractable solution to the Bayseisn estimation of non-Gaussian statistical models, an auxiliary function should 1) satisfy either the weak or strong condition and 2) have the same mathematical form as the prior distribution (up to a constant difference).

*1) Discussion:* Generally speaking, it is usually not feasible to find an auxiliary function that satisfies the strong condition, except that the original function $p(\mathbf{X},\mathbf{Z})$ is globally concave in terms of $\mathbf{Z}$ [3]. Unlike the case of the strong condition, it is relatively easy to find an auxiliary function to fulfill the weak condition, as the original function $p(\mathbf{X},\mathbf{Z})$ might be partially concave with respect to part of $\mathbf{Z}$ [5], [29]. For example [29], the multivariate log-inverse-beta (MLIB) function in the Dirichlet distribution is *not* globally concave in terms of all of its variables. It is only relatively concave *w.r.t.* one of its variables when fixing the rest. Iteratively taking this property, an auxiliary function that satisfies the weak condition (with a proper expectation form) and the requirement of the mathematical form can be found so that an analytically tractable solution is derived. In practice, an auxiliary function that satisfies either the strong or weak condition is difficult to design/obtain. One way of obtaining an appropriate auxiliary function is to consider the Jensen's inequality or the Taylor expansion, when combining with the convexity or relative convexity of the original function [2], [5].

In general, the weak condition yields smaller systematic gap in terms of approximation accuracy. Hence, if one can find an auxiliary function that satisfies the weak condition, there is no need to find another auxiliary function for the strong condition.

*2) Comparisons of Weak and Strong Conditions:* Since Dirichlet distribution is a multivariate case of beta distribution, the EVI-based Bayesian BMM that constructs an auxiliary function with the weak condition can be obtained based on the DMM work in [29] by simply setting the dimension to 2. The EVI-based Bayesian BMM proposed in [4] utilized the strong condition to choose the auxiliary function. Based on two different solutions for Bayesian estimation of BMM, we demonstrate the differences between the strong and weak conditions.

We consider the observation $x_n$ and the unobserved indicator vector $\mathbf{z}_n$ as the *complete* data. The conditional distribution of $\mathbf{X} = \{x_1,\ldots,x_N\}$ and $\mathbf{Z} = \{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$ given the parameters $\{\mathbf{U},\mathbf{V},\mathbf{\Pi}\}$ is

$$\begin{aligned}f(\mathbf{X},\mathbf{Z}|\mathbf{U},\mathbf{V},\mathbf{\Pi}) =& f(\mathbf{X}|\mathbf{U},\mathbf{V},\mathbf{Z}) f(\mathbf{Z}|\mathbf{\Pi}) \\ =& \prod_{n=1}^{N}\prod_{i=1}^{I}\left[\pi_i \mathrm{Beta}(x_n|u_i,v_i)\right]^{z_{ni}}.\end{aligned} \tag{15}$$

The ultimate goal is to estimate the posterior distributions of $u_i$, $v_i$, and $z_{ni}$, respectively.

In order to derive an analytically tractable solution to the posterior distributions, the most challenging part with the EVI framework is to calculate the expectation of the bivariate log-inverse-beta (LIB) function

$$\mathbf{E}_{u_i,v_i}\left[\mathrm{LIB}(u_i,v_i)\right] = \mathbf{E}_{u_i,v_i}\left[\frac{\Gamma(u_i+v_i)}{\Gamma(u_i)\Gamma(v_i)}\right]. \tag{16}$$

- *EVI-based Bayesian BMM with Weak Condition [29]*
  In the Bayesian BMM with SLB approximation [4], the new objective function that we are maximizing is

$$\begin{aligned}&\mathbf{E}_{\mathbf{Z}}\left[\ln \widetilde{p}_w(\mathbf{X},\mathbf{Z})\right] \\ =& \widetilde{\mathcal{L}}_{\mathrm{SLB}} \\ =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right]\left(\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i\right) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right]\left(\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i\right),\end{aligned} \tag{17}$$

  where $\overline{x}$ is the expected value of $x$ and $\psi(x)$ is the digamma function defined as $\psi(x) = \frac{\partial \ln\Gamma(x)}{\partial x}$. This lower-bound satisfies the weak condition such that $\mathbf{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X},\mathbf{Z})\right] \geq \mathbf{E}_{\mathbf{Z}}\left[\ln \widetilde{p}_w(\mathbf{X},\mathbf{Z})\right]$. Moreover, this lower-bound is identical for all the variables $u_i$, $v_i$, and $z_{ni}$
- *EVI-based Bayesian BMM with Strong Condition [4]*
  In the case of the strong condition, an auxiliary function $\widetilde{p}_s(\mathbf{X},\mathbf{Z})$ is required. In [4], three different auxiliary functions were derived for the variables $u_i$, $v_i$, and $z_{ni}$, respectively. To specify, for $u_i$, the auxiliary function is

$$\begin{aligned}\widetilde{p}_{s_{u_i}}(\mathbf{X},\mathbf{Z}) =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right]\left(\ln u_i-\ln\overline{u}_i\right) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right]\left(\ln v_i-\ln\overline{v}_i\right) \\ &+ \overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln u_i-\ln\overline{u}_i),\end{aligned} \tag{18}$$

[3] According to our experience, for (most of) the non-Gaussian statistical models, the original function is not globally concave.

[4] A Bayesian BMM with SLB approximation can be derived from the Bayesian DMM with SLB approximation [29] by setting the dimension of the Dirichlet variable to two.

where $\psi^{'}(x) = \frac{\partial \psi(x)}{\partial x}$. Hence, when considering $u_i$ as the variable, the objective function that is maximized is [4]

$$\begin{aligned}\widetilde{\mathcal{L}}_{\text{MLB}_{u_i}} =& \mathbf{E}_{\mathbf{Z}}\left[\widetilde{p}_{s_{u_i}}(\mathbf{X},\mathbf{Z})\right] \\ =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right](\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right](\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i) \\ &+ \overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i)+\text{const.}\end{aligned} \tag{19}$$

Similarly, due to the symmetry of $u_i$ and $v_i$, the objective function, when treating $v_i$ as the variable, is [4]

$$\begin{aligned}\widetilde{\mathcal{L}}_{\text{MLB}_{v_i}} =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right](\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right](\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i) \\ &+ \overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i)+\text{const.}\end{aligned} \tag{20}$$

When taking $z_{ni}$ as the only variable, the auxiliary function that proposed in [4] is

$$\begin{aligned}&\widetilde{p}_{s_{z_{ni}}}(\mathbf{X},\mathbf{Z}) \\ =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right](\ln u_i-\ln\overline{u}_i) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right](\ln v_i-\ln\overline{v}_i) \\ &+ 0.5\cdot\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right](\ln u_i-\ln\overline{u}_i)^2 \\ &+ 0.5\cdot\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right](\ln v_i-\ln\overline{v}_i)^2 \\ &+ \overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln u_i-\ln\overline{u}_i)(\ln v_i-\ln\overline{v}_i).\end{aligned} \tag{21}$$

Correspondingly, the objective function for updating the posterior distribution of $z_{ni}$ can be written as

$$\begin{aligned}&\widetilde{\mathcal{L}}_{\text{MLB}_{z_{ni}}} \\ =& \mathbf{E}_{\mathbf{Z}}\left[\widetilde{p}_{s_{z_{ni}}}(\mathbf{X},\mathbf{Z})\right] \\ =& \ln\frac{\Gamma(\overline{u}_i+\overline{v}_i)}{\Gamma(\overline{u}_i)\Gamma(\overline{v}_i)} \\ &+ \overline{u}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{u}_i)\right](\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i) \\ &+ \overline{v}_i\left[\psi(\overline{u}_i+\overline{v}_i)-\psi(\overline{v}_i)\right](\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i) \\ &+ 0.5\cdot\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right]\mathbf{E}\left[(\ln u_i-\ln\overline{u}_i)^2\right] \\ &+ 0.5\cdot\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right]\mathbf{E}\left[(\ln v_i-\ln\overline{v}_i)^2\right] \\ &+ \overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i)(\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i)\end{aligned}$$

It has been analyzed in Sec. III-B that both the strong condition and the weak condition incur systematic gaps. We now quantitatively compare the gaps. It is worth to note that the EVI-based Bayesian BMM with the strong condition is also a MLB approximation. In principle, there exist four combinations, which are "strong condition+SLB", "weak condition+SLB", "strong condition+MLB", and "weak condition+MLB". We focus only on the comparisons of weak and strong conditions in this section. The comparisons of the SLB approximation with the MLB approximation will be presented in the next section.

When taking $u_i$ as the variable, the difference between the objective functions obtained via weak and strong conditions, respectively, can be calculated as

$$\begin{aligned}\Delta\widetilde{\mathcal{L}}_{\text{SLB vs. MLB}_{u_i}} =& \widetilde{\mathcal{L}}_{\text{SLB}} - \widetilde{\mathcal{L}}_{\text{MLB}_{u_i}} \\ =& -\bar{u}_i\bar{v}_i\psi'(\bar{u}_i+\bar{v}_i)(\mathbf{E}\left[\ln u_i\right]-\ln\bar{u}_i) \\ \geq& 0,\end{aligned} \tag{22}$$

where we used the fact that $\psi^{'}(x) > 0$ and $\ln x$ is a convex function with respect to $x$. For $v_i$, it is straightforward to show that the difference is also positive by using the symmetric properties.

When comparing $\widetilde{\mathcal{L}}_{\text{SLB}}$ with $\widetilde{\mathcal{L}}_{\text{MLB}_{z_{ni}}}$, the difference is

$$\begin{aligned}&\Delta\widetilde{\mathcal{L}}_{\text{SLB vs. MLB}_{z_{ni}}} \\ =& \widetilde{\mathcal{L}}_{\text{SLB}} - \widetilde{\mathcal{L}}_{\text{MLB}_{z_{ni}}} \\ =& -\Big\{0.5\cdot\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right]\mathbf{E}\left[(\ln u_i-\ln\overline{u}_i)^2\right] \\ &+ 0.5\cdot\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right]\mathbf{E}\left[(\ln v_i-\ln\overline{v}_i)^2\right] \\ &+ \overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\mathbf{E}\left[\ln u_i\right]-\ln\overline{u}_i)(\mathbf{E}\left[\ln v_i\right]-\ln\overline{v}_i)\Big\}.\end{aligned}$$

It can be proved that the difference $\Delta\widetilde{\mathcal{L}}_{\text{SLB vs. MLB}_{z_{ni}}}$ is also greater than or equal to 0. More details for this proof can be found in Appendix A.

The aforementioned three positive differences indicate that the new objective function with the weak condition [29] is tighter (*i.e.*, closer to the original objective function) than that with the strong condition [4]. Thus, for the EVI-based Bayesian BMM, the systematic gap incurred by the weak condition is smaller than that incurred by the strong condition. This makes the weak condition more favorable in practice [5], [11], [29], [37]. Similar analysis can be applied to the Bayesian DMM with MLB [28] and the Bayesian DMM with SLB [29], as Dirichlet distribution is a multivariate extension of beta distribution.

### C. SLB Approximation and MLB Approximation

If we can find an auxiliary function $\widetilde{p}(\mathbf{X},\mathbf{Z})$ that contains all the variables $\mathbf{Z}$ and satisfies the aforementioned required conditions, the convergence of EVI is naturally guaranteed as this new objective function is convex or relatively convex in terms of $q_i(\mathbf{Z}_i)$ [17]. Since only one lower-bound approximation is applied to the original objective function, this approach is referred to as the single lower-bound (SLB) approximation and has been applied in, *e.g.*, [5], [11], [29].

When dividing $\mathbf{Z}$ into $M$ disjoint groups as $\mathbf{Z} = \{\mathbf{Z}_1,\ldots,\mathbf{Z}_i,\ldots,\mathbf{Z}_M\}$, there might exist several auxiliary

TABLE II
SLB AND MLB APPROXIMATION COMPARISONS WITH DIFFERENT NON-GAUSSIAN STATISTICAL MODELS.

(a) Comparisons of the objective functions for Bayesian BMM.

| Model | Parameters | $\mathcal{L}_{\text{SLB}} - \mathcal{L}_{\text{MLB}}$ | $\text{KL}_{\text{SLB}} - \text{KL}_{\text{MLB}}$ |
|---|---|---|---|
| A | $\pi_1 = 0.3, u_1 = 2, v_1 = 8$<br>$\pi_2 = 0.7, u_2 = 15, v_2 = 4$ | $3.6 \times 10^{-3}$ | $-2.8 \times 10^{-3}$ |
| B | $\pi_1 = 0.3, u_1 = 10, v_1 = 2$<br>$\pi_2 = 0.4, u_2 = 2, v_2 = 12$<br>$\pi_3 = 0.3, u_3 = 10, v_3 = 10$ | $1.3 \times 10^{-3}$ | $-0.58 \times 10^{-3}$ |

(b) Comparisons of the objective functions for Bayesian DMM.

| Model | Parameters | $\mathcal{L}_{\text{SLB}} - \mathcal{L}_{\text{MLB}}$ | $\text{KL}_{\text{SLB}} - \text{KL}_{\text{MLB}}$ |
|---|---|---|---|
| A | $\pi_1 = 0.65,\ \mathbf{u}_1 = [4\ 12\ 3]^{\text{T}}$<br>$\pi_2 = 0.35,\ \mathbf{u}_2 = [10\ 6\ 2]^{\text{T}}$ | $2.6 \times 10^{-3}$ | $-1.8 \times 10^{-3}$ |
| B | $\pi_1 = 0.2,\ \mathbf{u}_1 = [3\ 5\ 12\ 6]^{\text{T}}$<br>$\pi_2 = 0.5,\ \mathbf{u}_2 = [4\ 12\ 3\ 9]^{\text{T}}$<br>$\pi_3 = 0.3,\ \mathbf{u}_3 = [10\ 6\ 2\ 5]^{\text{T}}$ | $4.5 \times 10^{-3}$ | $-4.2 \times 10^{-3}$ |

(c) Comparisons of the objective functions for Bayesian GiDMM.

| Model | Parameters | $\mathcal{L}_{\text{SLB}} - \mathcal{L}_{\text{MLB}}$ | $\text{KL}_{\text{SLB}} - \text{KL}_{\text{MLB}}$ |
|---|---|---|---|
| A | $\pi_1 = 0.3,\ \mathbf{u}_1 = [3\ 10\ 5]^{\text{T}},\ \mathbf{v}_1 = [8\ 10\ 4]^{\text{T}}$<br>$\pi_2 = 0.4,\ \mathbf{u}_2 = [12\ 3\ 8]^{\text{T}},\ \mathbf{v}_2 = [4\ 9\ 4]^{\text{T}}$<br>$\pi_3 = 0.3,\ \mathbf{u}_3 = [12\ 4\ 7]^{\text{T}},\ \mathbf{v}_3 = [6\ 8\ 10]^{\text{T}}$ | 19.55 | $-5.6 \times 10^{-3}$ |
| B | $\pi_1 = 0.2,\ \mathbf{u}_1 = [25\ 15\ 40\ 32\ 5]^{\text{T}},\ \mathbf{v}_1 = [10\ 28\ 21\ 12\ 10]^{\text{T}}$<br>$\pi_2 = 0.4,\ \mathbf{u}_2 = [18\ 10\ 5\ 25\ 40]^{\text{T}},\ \mathbf{v}_2 = [24\ 18\ 20\ 8\ 16]^{\text{T}}$<br>$\pi_3 = 0.3,\ \mathbf{u}_3 = [48\ 28\ 15\ 8\ 36]^{\text{T}},\ \mathbf{v}_3 = [10\ 16\ 10\ 20\ 18]^{\text{T}}$ | 10.15 | $-7.2 \times 10^{-3}$ |

functions. For example, we could have $M$ auxiliary functions as

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}) \geq & \widetilde{p}_1(\mathbf{X}, \mathbf{Z}_1) \\ & \cdot \\ p(\mathbf{X}, \mathbf{Z}) \geq & \widetilde{p}_i(\mathbf{X}, \mathbf{Z}_i) \\ & \cdot \\ p(\mathbf{X}, \mathbf{Z}) \geq & \widetilde{p}_M(\mathbf{X}, \mathbf{Z}_M). \end{aligned} \tag{23}$$

This approach is referred to as the multiple lower-bound (MLB) approximation. As each of the above mentioned auxiliary functions satisfies the required conditions in Sec. III-B, the optimal solution in (6) is

$$\ln \widetilde{q}_i^*(\mathbf{Z}_i) = \mathbf{E}_{\mathbf{Z}} \left[\ln \widetilde{p}_i(\mathbf{X}, \mathbf{Z}_i)\right] + \text{const} = \ln \widetilde{p}_i(\mathbf{X}, \mathbf{Z}_i) + \text{const}. \tag{24}$$

In this case, the new objective function that is maximized during each iteration is *not unique*. Hence, *there is no globally objective function that is maximized during each iteration*. Thus, the convergence cannot be theoretically guaranteed. This approach has been applied in [4] and [28]. Although convergence is not theoretically guaranteed, it can be monitored empirically.

Let us study a simple case with two disjoint groups in the MLB approximation. Assuming that $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2\}$ and we have two auxiliary functions $\widetilde{p}_1(\mathbf{X}, \mathbf{Z}_1)$ and $\widetilde{p}_2(\mathbf{X}, \mathbf{Z}_2)$ for $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively. As mentioned above, two different lower-bounds are obtained as

$$\begin{aligned} \widetilde{\mathcal{L}}_1 =& \mathbf{E}_{\mathbf{Z}} \left[\ln \widetilde{p}_1(\mathbf{X}, \mathbf{Z}_1) - \ln q(\mathbf{Z})\right] \\ \widetilde{\mathcal{L}}_2 =& \mathbf{E}_{\mathbf{Z}} \left[\ln \widetilde{p}_2(\mathbf{X}, \mathbf{Z}_2) - \ln q(\mathbf{Z})\right]. \end{aligned} \tag{25}$$

If we maximize each lower-bound separately, the optimal solutions to these two disjoint groups are

$$\ln \widetilde{q}_1^*(\mathbf{Z}_1) = \mathbf{E}_{\mathbf{Z} \backslash \mathbf{Z}_1} \left[\ln \widetilde{p}_1(\mathbf{X}, \mathbf{Z}_1)\right] + \text{const} \tag{26a}$$

$$\ln \widetilde{q}_2^*(\mathbf{Z}_2) = \mathbf{E}_{\mathbf{Z} \backslash \mathbf{Z}_2} \left[\ln \widetilde{p}_2(\mathbf{X}, \mathbf{Z}_2)\right] + \text{const}. \tag{26b}$$

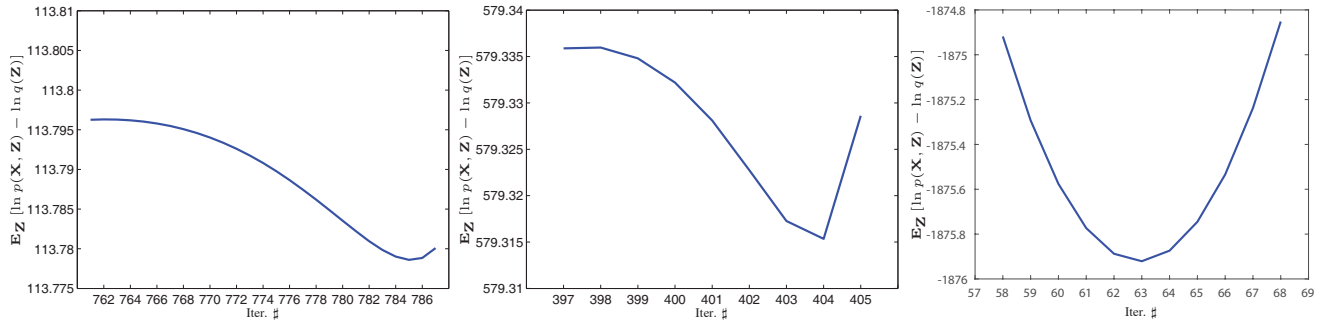
With these solutions, it appears what we are maximizing is just two times of the original lower-bound as

$$2 \times \mathcal{L} \geq \widetilde{\mathcal{L}}_1 + \widetilde{\mathcal{L}}_2 \tag{27a}$$

$$= \mathbf{E}_{\mathbf{Z}} \left[\ln \widetilde{p}_1(\mathbf{X}, \mathbf{Z}_1)\right] - \mathbf{E}_{\mathbf{Z}} \left[\ln q(\mathbf{Z})\right] \tag{27b}$$

$$+ \mathbf{E}_{\mathbf{Z}} \left[\ln \widetilde{p}_2(\mathbf{X}, \mathbf{Z}_2)\right] - \mathbf{E}_{\mathbf{Z}} \left[\ln q(\mathbf{Z})\right]. \tag{27c}$$

When performing the update strategy (26a), we get (27b) to be maximized. It is due to the fact that the optimal solution $\ln \widetilde{q}_1^*(\mathbf{Z}_1)$ maximizes $\widetilde{\mathcal{L}}_1$. This maximization makes the distribution of $\mathbf{Z}_1$ to be less uncertain. As $-\mathbf{E}_{\mathbf{Z}} \left[\ln q(\mathbf{Z})\right]$ in (27c) is the differential entropy of $\mathbf{Z}$, (27c) is decreasing while (27b) is maximizing. It is hard to evaluate if (27b) changes more than (27c) or not. Thus, the overall lower-bound, *i.e.*, $\widetilde{\mathcal{L}}_1 + \widetilde{\mathcal{L}}_2$ in (27a), might decrease during some iterations. On the one hand, as the lower-bound (*i.e.*, $\widetilde{\mathcal{L}}_1 + \widetilde{\mathcal{L}}_2$) to the original objective function cannot be guaranteed to be maximized all the time, this strategy may not promise convergence. On the other hand, if the change in (27b) is larger than that in (27c), the convergence is still guaranteed. There is no general judgement for the convergence. It should be studied case by case. Similar arguments can be applied to the case with more than two auxiliary functions. Thus, the convergence of MLB approximation is unguaranteed. In summary, SLB approximation can theoretically guarantee the




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

Fig. 3. Observations of decreasing values of the objective function during iterations. In principle, objective function should always increase (at least not decrease). This non-convergence fact indicates that the MLB approximation-based method may not promise convergence. Model A is a BMM with parameter $\pi_1 = 0.3, \pi_2 = 0.7, \mathbf{u}_1 = [2\ 8]^{\mathrm{T}}, \mathbf{u}_2 = [15\ 4]^{\mathrm{T}}$, model B is a three-dimensional DMM with parameter $\pi_1 = 0.35, \pi_2 = 0.65, \mathbf{u}_1 = [4\ 12\ 3]^{\mathrm{T}}, \mathbf{u}_2 = [10\ 6\ 2]^{\mathrm{T}}$, and model C is a three-dimensional GiDMM with parameter $\pi_1 = 0.3, \pi_2 = 0.4, \pi_3 = 0.3, \mathbf{u}_1 = [3\ 10\ 5]^{\mathrm{T}}$, $\mathbf{v}_1 = [8\ 10\ 4]^{\mathrm{T}}$, $\mathbf{u}_2 = [12\ 3\ 8]^{\mathrm{T}}$, $\mathbf{v}_2 = [4\ 9\ 4]^{\mathrm{T}}$, $\mathbf{u}_3 = [12\ 4\ 7]^{\mathrm{T}}$, $\mathbf{v}_3 = [6\ 8\ 10]^{\mathrm{T}}$. 400 samples were generated from each model.

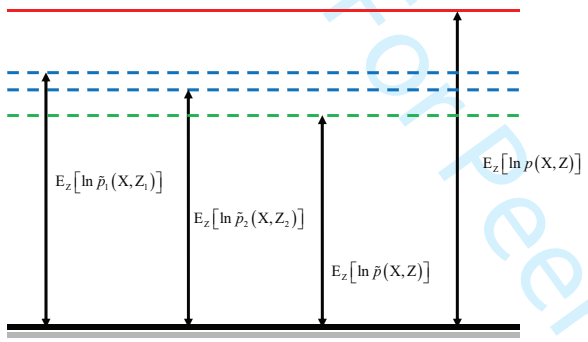



Fig. 2. Qualitative comparisons of SLB and MLB. For MLB, two different lower-bounds are introduced for $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively (the blue lines). For SLB, there is only one lower-bound (the green line). The original objective function is marked with red solid line. It can be observed that the new objective function that needs to be maximized is not unique in the MLB case. Hence, the convergence is not guaranteed. A new single objective function is employed and maximized in the SLB case. Therefore, the convergence is theoretically guaranteed.

convergence while MLB approximation, in general, cannot promise convergence.[5]

*1) Comparisons of MLB and SLB Approximations:* In the previous section, we analyzed and compared the weak and strong conditions for the EVI framework. Another important issue in EVI implementation is to distinguish the MLB and SLB approximations, as the latter can guarantee convergence but the former may not. To this end, we compare the MLB approximation-based algorithm with the SLB approximation-based algorithm in this section.

- *Observations of Oscillation*

  As discussed in Sec. III-C, the convergence of the MLB method is not guaranteed. We ran the MLB approximation-based Bayesian BMM algorithm [4], Bayesian DMM algorithm [28], and Bayesian GiDMM [48], respectively, and monitored the value of the objective function during each iteration. It was observed that, for some rounds of simulations [6], the objective function was *oscillating* during some iterations. This phenomenon was observed for several times, for BMM, DMM, and GiDMM. Figure 3 illustrates the decreasing objective function values and the corresponding iterations. For the SLB approximation-based Bayesian BMM and Bayesian DMM [29], the monitored objective function was always increasing until convergence. The observations of oscillation demonstrate that the convergence with MLB approximation cannot be guaranteed.

- *Comparisons of Estimation Accuracy*

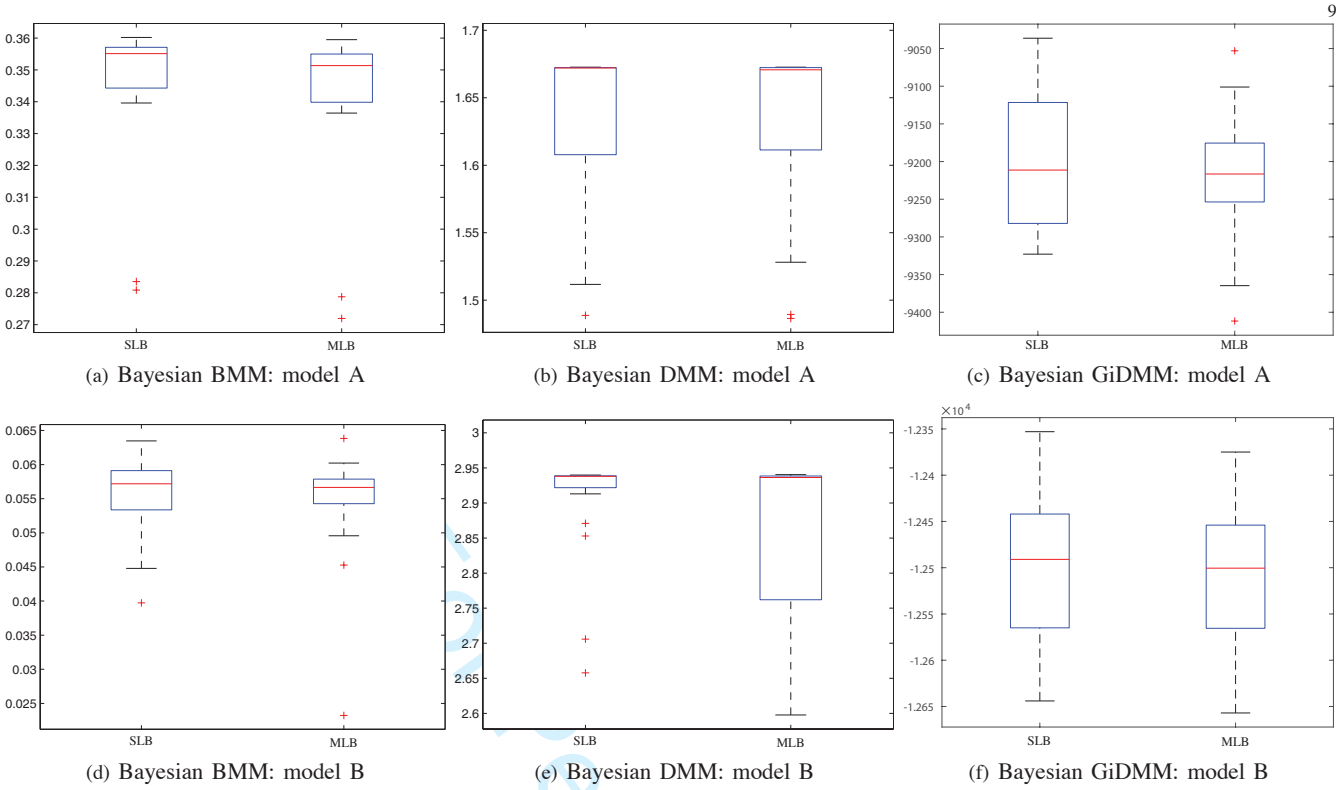
  In this section, we compare the MLB approximation with the SLB approximation quantitatively. With a known BMM or DMM or GiDMM, $2,000$ samples were generated, respectively. The above-mentioned Bayesian estimation algorithms were applied to estimate the posterior distributions, respectively. We calculated the original variational objective function in (1) to examine which approximation is better. With the obtained posterior distribution $q^*(\mathbf{Z})$, the original variational objective function is calculated numerically by sampling methods. Hence,we got two different values, $\mathcal{L}_{\mathrm{SLB}}$ and $\mathcal{L}_{\mathrm{MLB}}$, from the SLB approximation and the MLB approximation, respectively. Larger value means closer lower-bound approximation. In addition to this, we also measure the estimation accuracy by the KL divergence of the estimated PDF from the true one as $\mathrm{KL}(p(\mathbf{X}|\mathbf{\Theta})\|p(\mathbf{X}|\widehat{\mathbf{\Theta}}))$, where $\mathbf{\Theta}$ is the true parameter vector and $\widehat{\mathbf{\Theta}}$ is the estimated one. Similarly, we numerically calculated $\mathrm{KL}_{\mathrm{SLB}}$ and $\mathrm{KL}_{\mathrm{MLB}}$ from the SLB and MLB approximations [7], respectively. The smaller the KL divergence is, the more accurate the estimation is.

  For Bayesian BMM, comparisons are presented in Table II(a), Figure 4(a) and 4(d). The comparisons of the Bayesian DMM via SLB [29] and MLB [28] approximations are illustrated in Table II(b), Figure 4(b) and 4(e). For GiDMM, the comparisons of Bayesian GiDMM via SLB (algorithm presented in Appendix B) and MLB [48]

[5] In practice (*e.g.*, [4], [28]), the EVI-based algorithm may also converge with MLB approximation. However, it is empirical result without proof.

[6] Here, one simulation round means that we ran the estimation algorithm until it stops according to some criterion.

[7] For the MLB approximation, we only take those simulation rounds that always converge into consideration.




(a) Bayesian BMM: model A (b) Bayesian DMM: model A (c) Bayesian GiDMM: model A

(d) Bayesian BMM: model B (e) Bayesian DMM: model B (f) Bayesian GiDMM: model B

Fig. 4. Comparisons of the original objective functions for Bayesian BMM, Bayesian DMM, and Bayesian GiDMM (with SLB and MLB, respectively). For each sub-figure, 20 rounds of simulations were conducted. In each boxplot, mark is the median, the box edges are the $25^{th}$ and $75^{th}$ percentiles. The outliers are marked individually. Model settings are the same as Table II.

approximations are shown in Table II(c), Figure 4(c) and 4(f). All the simulations were run 20 rounds and the mean values are reported.

It can be observed that, for Bayesian BMM, Bayesian DMM and Bayesian GiDMM, the SLB approximation yields higher objective function value than the MLB approximation, respectively. Meanwhile, the KL divergences obtained by the SLB approximation are all smaller than those obtained by the MLB. The results suggest that the SLB approximation is superior to the MLB approximation.

## IV. Experimental Results and Discussions

In this section, we compare the SLB-based GiDMM with the MLB-based GiDMM using the tasks of text categorization, object detection, and image categorization.

We apply several statistical models to describe the underlying distribution of the structured features extracted from text and images. It is worth noting that the main motivation of the real data evaluation is to evaluate and analyze the multiple/single lower bound approximations for non-Gaussian mixture models. Hence, we do not involve other non-mixture model-based methods for comparisons.

*2) Text Categorization:* With the development of internet, the amount of the text documents has been dramatically increased. If the text documents are organized and processed manually, it not only will consume a great deal of manpower, but also is hard to conduct efficient analysis [49], [50]. Hence, there exists urgent need to develop a technique which can organize the documents efficiently. In order to meet the requirements and challenges, automatic text categorization [50]–[52], which is a key technique for processing and organizing vast amount of text data, has been widely applied. It has very realistic significance for efficient management and effective utilization of information and has gradually become an important research direction in the domain of data mining [47]. Recently, different statistical model-based approaches have been proposed and utilized to carry out the text categorization task [30], [51]–[55].

In this paper, we report the experimental results by using the Bayesian GiDMM as a classifier for the task of text categorization on the dataset gathered from the top 10 largest categories of the "ModApte" split of the Reuters-21578[8]. This dataset is composed of 9990 news stories which were grouped into 10 categorizations. Each categorization is randomly split into two halves, one half for training and the other half for test. Following the work in [56], the Porter's stemming [57] is used to reduce the words to their base forms. In this pre-processing stage, the words that occur less than 3 times or are shorter than 2 in length are eliminated. Eventually, each document was represented by a 10-dimensional vector which contains only positive elements. Based on the aforementioned pre-processing, we trained a statistical model (Bayesian GiDMM with SLB approximation, denoted as "Bayesian GiDMM$_{\text{SLB}}$". Detailed algorithm is presented in Appendix B) for each categorization and calculate the likelihood for the test vectors. A test document is considered correctly categorized if its corresponding model yields the highest likelihood. For comparison, we have also applied four other approaches for categorizing text documents: the Bayesian GiDMM using the MLB

[8]http://kdd.ics.uci.edu/databases/reuters21578/

TABLE III
TEXT CATEGORIZATION ACCURACIES OBTAINED BY DIFFERENT METHODS.

| Method | Bayesian GiDMM$_{\text{SLB}}$ | Bayesian GiDMM$_{\text{MLB}}$ | Bayesian iDMM$_{\text{SLB}}$ | Bayesian iDMM$_{\text{MLB}}$ | Bayesian GMM |
|---|---|---|---|---|---|
| Accuracy (in %) | **86.89** | 86.45 | 86.02 | 85.82 | 80.63 |
| Runtime (in s)$^{\dagger}$ | **5.33** | 11.71 | 8.20 | 13.56 | 12.41 |

$^{\dagger}$ On a ThinkCentre® computer with Intel® Core™ i5 − 4590 CPU 4G.

approximation (denoted as "Bayesian GiDMM$_{\text{MLB}}$" [48].), the iDMM using the SLB and MLB approximation (which we refer to as "Bayesian iDMM$_{\text{SLB}}$" [58] and "Bayesian iDMM$_{\text{MLB}}$" [59], respectively), and the Bayesian Gaussian mixture model (Bayesian GMM) [17]. The main motivation is to validate the approaches of text categorization by considering comparable statistical model-based methods. Table III shows the categorization accuracies. It can be observed that the best performance is obtained by "Bayesian GiDMM$_{\text{SLB}}$", in terms of the categorization accuracy rate (*i.e.*, 86.89%), which demonstrates the advantage of using the SLB approximation over using the MLB approximation, for the non-Gaussian statistical model-based text categorization task. The reported values are the means of 10 times evaluations. In each evaluation, the aforementioned procedures are repeated.

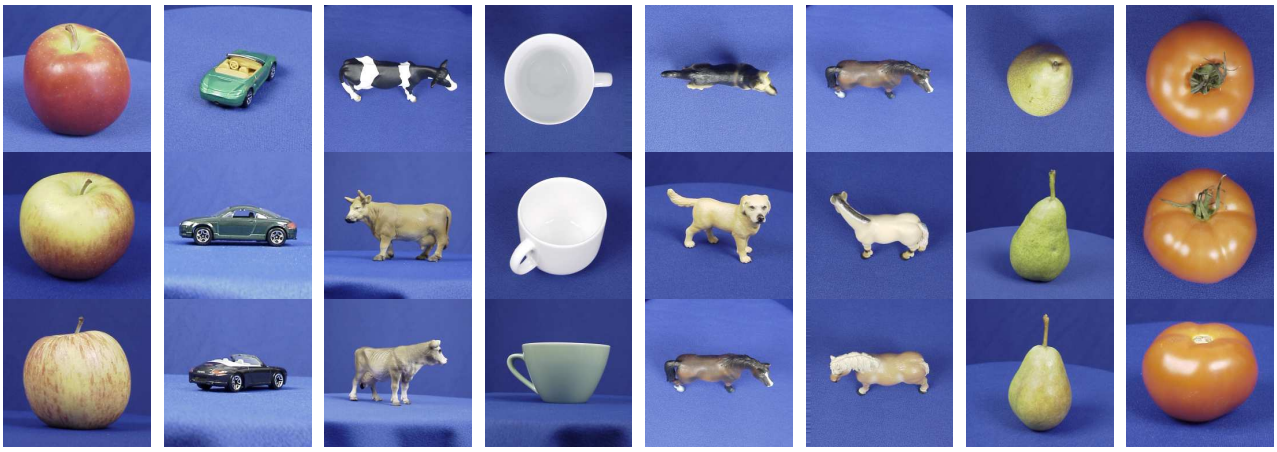
*3) Object Detection:* Object detection refers to the task of distinguishing a specific object (*e.g.*, car, face) from other objects in an image, which is an essential task in computer vision. It has been a topic of extensive studies in the past decades. Object detection has various applications, such as robotics [60], medical image analysis [61], surveillance and human computer interaction [62]. Although humans usually perform well in object detection, it is much difficult to obtain similar performance for the machines, which is mainly due to the changes in illumination conditions, orientations, positions and scales. The aforementioned factors can dramatically affect the appearance of a given object. Recently, a great deal of research efforts have devoted to overcome such difficulties [33], [63]–[65]. These researches can be divided into two main categories. The first one has been devoted to the development of excellent global or local visual image descriptors [66]. The second one has been focusing on the development of powerful and robust classifiers [67].

As with the majority of computer vision tasks, a key step for accurate object detection is to extract good descriptors to represent these target objects. Recently, researchers have proposed many global and local visual descriptors, such as the Histogram of Oriented Gradient (HOG) descriptor [68], which is originally developed for detecting pedestrian in gray-scale images. Here, we use the rectangular HOG (denoted as RHOG) descriptor [69], which generates positive feature vectors. Moreover, it is found to be efficient and convenient for our object detection task. Experiments were conducted by considering seven windows for the RHOG descriptor, such that each image can be represented by a 441-dimensional feature vector. We use the publicly available ETH-80 dataset [70]. This dataset consists of 3280 images, which are categorized into eight object classes. Each class contains 10 unique objects and 41 views (*i.e.*, 410 images for each class). Example images from this dataset are shown in Figure 5. During the evaluation, we trained one detector for each class. For a given class, we take all the images from this class as the positive set and the images in the negative set for this class were taken from the other seven remaining classes, where each class contributes $1/7$ (approx. 58 images from each of the seven classes) to the negative examples. The detector was trained on half of the positive set and half of the negative set, where these two sets were randomly split into two equal parts, respectively. The remaining half of the positive and negative sets were used as the test set.

With the above training/test set selection, our methodology for object detection can be summarized as follows. First, RHOG descriptors were extracted from each image. By doing this, the description of each image was represented as a positive vector. Second, each vector is assumed generated from a mixture of generalized inverted-Dirichlet distributions and we apply the proposed Bayesian GiDMM$_{\text{SLB}}$ as a classifier to detect objects by assigning the test image to the group which has the highest posterior probability according to Bayes' decision rule. Similar to the text categorization task in Sec. IV-2, we also train classifiers based on Bayesian GiDMM$_{\text{MLB}}$, Bayesian iDMM$_{\text{SLB}}$, Bayesian iDMM$_{\text{MLB}}$, and Bayesian GMM, respectively. We report on the accuracies of the aforementioned five classifiers in Table IV. As can be observed from this table, the Bayesian GiDMM$_{\text{SLB}}$ achieves the best detection rates, compared to the other referred methods. Similar to the task of text categorization, it is also observed that the SLB approximation is superior to the MLB approximation for both Bayesian GiDMM and Bayesian iDMM in the object detection task. We conducted 10 rounds of simulations and the mean values are reported.

*4) Image Categorization:* With the development and broad applications of digital information acquiring techniques, the number of digital images has grown enormously. Image categorization task is developed to meet the requirements of many important applications, such as image retrieval [71], content-based images recommendation [72], and automatic image understanding [73]. Recently, image categorization has emerged as an attractive area in computer vision [66], [74], [75]. A key step for accurate images categorization is to extract robust and efficient image descriptors to represent these images. Here, we use the RHOG descriptor [69] again by considering eight windows for the RHOG descriptor such that each image in the dataset was then represented by a 576-dimensional feature vector.

The evaluations were based on the MIT Scene dataset [76] which is composed of 2688 images categorized into eight categories. The categories are coast (360 images), forest (328 images), mountain (374 images), open country (410 images), highway (260 images), inside of cities (308 images), tall building (356 images), and street (292 images). All of the color images are in JPEG format, and the size of each image is $256 \times 256$. A few example images from this dataset are shown in Figure 6. This dataset can be divided into two subsets: the



(a) Apple. (b) Car. (c) Cow. (d) Cup. (e) Dog. (f) Horse. (g) Pear. (h) Tomato.

Fig. 5. Example images from the ETH-80 dataset. Each column represent a set of examples from one class.

TABLE IV
THE AVERAGE DETECTION RATES (%) ON THE ETH-80 DATASET, USING DIFFERENT METHODS.

| Class | Bayesian GiDMM$_{SLB}$ | Bayesian GiDMM$_{MLB}$ | Bayesian iDMM$_{SLB}$ | Bayesian iDMM$_{MLB}$ | Bayesian GMM |
|---|---|---|---|---|---|
| Apple | **96.96** | 95.76 | 92.26 | 91.51 | 88.33 |
| Car | **97.21** | 95.78 | 89.26 | 88.34 | 87.16 |
| Cow | **88.73** | 86.94 | 83.88 | 82.43 | 80.95 |
| Cup | **99.51** | 98.39 | 92.12 | 91.64 | 87.34 |
| Dog | **87.89** | 84.73 | 82.58 | 81.17 | 80.35 |
| Horse | **87.81** | 86.79 | 80.24 | 78.15 | 76.82 |
| Pear | **99.39** | 98.95 | 95.76 | 93.89 | 90.14 |
| Tomato | **98.73** | 97.76 | 93.37 | 92.69 | 89.37 |

natural subset and the man-made subset. The natural subset contains 1472 images from four different categories: coast, forest, mountain, and open country. The man-made subset has 1216 images from another four different categories: highway, inside of cities, tall building, and street. During evaluations, each category was randomly split into two separate halves, one for training and the other for test. For each category, the feature vectors in the training sets were then modeled by the Bayesian GiDMM$_{SLB}$. Finally, the Bayes classification rule was applied to assign each test vector to a given class according to their posterior probabilities. With the same procedure mentioned above, four other statistical models, Bayesian GiDMM$_{MLB}$, Bayesian iDMM$_{SLB}$, Bayesian iDMM$_{MLB}$, and Bayesian GMM were also applied. Ten rounds of simulations were conducted and we reported the mean value in Table V. It is clearly shown that the Bayesian GiDMM$_{SLB}$ achieved the highest accuracy rate, compared with other methods. It is also observed that the SLB approximation-based methods outperforms the MLB approximation-based methods.

## V. CONCLUSIONS

Structured data are ubiquitous existing in daily life. Compared with the conventional Gaussian distributed data, such type of data have different properties and distributions. Hence, specific non-Gaussian statistical models are required to applied. The extended variational inference (EVI) framework can be efficiently implemented in estimation of non-Gaussian statistical models. We discussed and summarized the required conditions for selection of the auxiliary functions in the EVI framework. Moreover, we also analyzed and compared the multiple lower-bounds (MLB) approximation and the single lower-bound (SLB) approximation. Theoretical analysis showed that the weak condition, in general, incurs smaller systematic gap than the strong condition. Hence, if the auxiliary function under the weak condition can be obtained, the weak condition is preferable. Otherwise, we can apply either the strong or weak condition to design/obtain an auxiliary function to carry out EVI. Synthesized structured data evaluations with Bayesian beta mixture model, Bayesian Dirichlet mixture model, and Bayesian generalized inverted-Dirichlet mixture model demonstrated that the SLB approximation can theoretically guarantee convergence and is superior to the MLB approximation. The advantages of the SLB approximation over the MLB approximation were also illustrated by three real-life structured data-based applications.

## ACKNOWLEDGEMENT

This work was partly supported by the National Key R&D Program of China No. 2018YFC0807205, by the National Natural Science Foundation of China No. 61773071, 61628301, by the Beijing Nova Program No. Z171100001117049, by the Beijing Nova Program Interdisciplinary Cooperation Project No. Z181100006218137, by the National Science and Technology Major Program of the Ministry of Science and Technology No.2018ZX03001031, by the Key program of Beijing Municipal Natural Science Foundation No. L172030.

(a) Coast. (b) Forest. (c) Mountain. (d) Open country. (e) Highway. (f) Incidecity. (g) Street. (h) Tall building.

Fig. 6. Example images from MIT Scene dataset [76].

TABLE V
THE AVERAGE CATEGORIZATION ACCURACIES (IN %) ON MIT SCENE DATASET, USING DIFFERENT METHODS.

| Category | Bayesian GiDMM$_{\text{SLB}}$ | Bayesian GiDMM$_{\text{MLB}}$ | Bayesian iDMM$_{\text{SLB}}$ | Bayesian iDMM$_{\text{MLB}}$ | Bayesian GMM |
|---|---|---|---|---|---|
| All | **73.04** | 72.05 | 68.14 | 67.88 | 65.17 |
| Natural | **76.29** | 75.30 | 73.11 | 72.58 | 70.76 |
| Man-made | **79.12** | 77.83 | 74.63 | 74.27 | 71.12 |

## APPENDIX A
## PROOF OF $\Delta\widetilde{\mathcal{L}}_{\text{SLB vs. MLB}_{z_{ni}}} \geq 0$

Denoting

$$\begin{aligned}\mathbf{G}(u_i,v_i) =& 0.5\cdot\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right](\ln u_i-\ln\overline{u}_i)^2\\ &+0.5\cdot\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right](\ln v_i-\ln\overline{v}_i)^2\\ &+\overline{u}_i\cdot\overline{v}_i\cdot\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln u_i-\ln\overline{u}_i)(\ln v_i-\ln\overline{v}_i).\end{aligned} \tag{28}$$

The Hessian of $\mathbf{G}(u_i,v_i)$ with respect to $[\ln u_i,\ln v_i]^{\text{T}}$ is

$$\nabla^2\mathbf{G}(u_i,v_i)=\begin{bmatrix}\frac{\partial^2\mathbf{G}(u_i,v_i)}{(\partial\ln u_i)^2} & \frac{\partial^2\mathbf{G}(u_i,v_i)}{\partial\ln u_i\partial\ln v_i}\\ \frac{\partial^2\mathbf{G}(u_i,v_i)}{\partial\ln v_i\partial\ln u_i} & \frac{\partial^2\mathbf{G}(u_i,v_i)}{(\partial\ln v_i)^2}\end{bmatrix}, \tag{29}$$

where

$$\begin{aligned}\frac{\partial^2\mathbf{G}(u_i,v_i)}{(\partial\ln u_i)^2}=&\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right],\\ \frac{\partial^2\mathbf{G}(u_i,v_i)}{\partial\ln u_i\partial\ln v_i}=&\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i),\\ \frac{\partial^2\mathbf{G}(u_i,v_i)}{\partial\ln v_i\partial\ln u_i}=&\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i),\\ \frac{\partial^2\mathbf{G}(u_i,v_i)}{(\partial\ln v_i)^2}=&\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right].\end{aligned} \tag{30}$$

The determinant of $\nabla^2\mathbf{G}(u_i,v_i)$ is

$$\begin{aligned}&|\nabla^2\mathbf{G}(u_i,v_i)|\\ =&\overline{u}_i^2\overline{v}_i^2\left\{\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right]\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right]\right.\\ &\left.-\left[\psi^{'}(\overline{u}_i+\overline{v}_i)\right]^2\right\}.\end{aligned} \tag{31}$$

Since $\psi^{'}(x)$ is a monotonously decreasing function and $\lim_{x\to\infty}\psi^{'}(x)=0$, we have

$$\begin{aligned}\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)<&\psi^{'}(\overline{u}_i+\overline{v}_i)\\ \psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)<&\psi^{'}(\overline{u}_i+\overline{v}_i).\end{aligned} \tag{32}$$

Hence, $|\nabla^2\mathbf{G}(u_i,v_i)|<0$ and $\mathbf{G}(u_i,v_i)$ is a concave function in terms of $[\ln u_i,\ln v_i]^{\text{T}}$.

In order to calculate the maximum value of $\mathbf{G}(u_i,v_i)$, we have

$$\begin{aligned}\frac{\partial G(u_i,v_i)}{\partial\ln u_i}=&\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right](\ln u_i-\ln\overline{u}_i)\\ &+\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln v_i-\ln\overline{v}_i)=0,\\ \frac{\partial G(u_i,v_i)}{\partial\ln v_i}=&\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right](\ln v_i-\ln\overline{v}_i)\\ &+\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln u_i-\ln\overline{u}_i)=0.\end{aligned}$$

Then, we get

$$\begin{aligned}&\overline{u}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{u}_i)\right](\ln u_i-\ln\overline{u}_i)\\ =&-\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln v_i-\ln\overline{v}_i),\end{aligned} \tag{33}$$

$$\begin{aligned}&\overline{v}_i^2\left[\psi^{'}(\overline{u}_i+\overline{v}_i)-\psi^{'}(\overline{v}_i)\right](\ln v_i-\ln\overline{v}_i)\\ =&-\overline{u}_i\overline{v}_i\psi^{'}(\overline{u}_i+\overline{v}_i)(\ln u_i-\ln\overline{u}_i).\end{aligned} \tag{34}$$

Substituting (33) and (34) into (28) and with some algebra, we get the maximum value of $\mathbf{G}(u_i,v_i)$ as

$$\max_{u_i,v_i}\mathbf{G}(u_i,v_i)=0. \tag{35}$$

Finally, we can conclude that $\mathbf{G}(u_i,v_i)\leq 0$ and $\Delta\widetilde{\mathcal{L}}_{\text{SLB vs. MLB}_{z_{ni}}}=-\mathbf{E}\left[\mathbf{G}(u_i,v_i)\right]\geq 0$.

## APPENDIX B
## ALGORITHM FOR BAYESIAN GIDMM WITH SLB

Following the approach proposed in [48], the estimation of the parameters in (17) is equivalent to the estimation of the parameters in the following mixture model

$$f(\mathbf{Y},\mathbf{\Pi},\mathbf{U},\mathbf{V})=\prod_{n=1}^{N}\sum_{i=1}^{I}\pi_i\prod_{k=1}^{K}\text{iBeta}(y_{nk};u_{ik},v_{ik}), \quad (36)$$

where $y_{n1}=x_{n1}$ and $y_{nk}=x_{nk}/(1+\sum_{l=1}^{K-1}x_{nl})$ for $k>1$ and $\text{iBeta}(y;u,v)$ is an inverted Beta distribution defined with parameters $(u,v)$, defined as

$$\text{iBeta}(y;u,v)=\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}y^{u-1}(1+y)^{-u-v},\ x>0. \quad (37)$$

For each observation $\mathbf{x}_n$ (also for $\mathbf{y}_n$), we introduce an $I$-dimensional binary random vector $\mathbf{z}_n=(z_{n1},\cdots,z_{nI})$, specifying which component that $\mathbf{x}_n$ belongs to. If $\mathbf{x}_n$ is generated from component $i$, $z_{ni}=1$; otherwise, $z_{ni}=0$. The prior distribution of $\mathbf{Z}$, given the mixing coefficients $\mathbf{\Pi}$, is defined as

$$p(\mathbf{Z}|\mathbf{\Pi})=\prod_{n=1}^{N}\prod_{i=1}^{I}\pi_i^{z_{ni}}. \quad (38)$$

To perform the variational inference of the GiDMM, we have to place conjugate priors over the model parameters $\mathbf{U}$ and $\mathbf{V}$. Here, we consider the Gamma distribution as a conjugate prior distribution for them as

$$\begin{aligned}f(\mathbf{U})&=\prod_{i=1}^{I}\prod_{k=1}^{K}\text{Gam}(u_{ik};g_{ik},h_{ik})\\&=\prod_{i=1}^{I}\prod_{k=1}^{K}\frac{h_{ik}^{g_{ik}}}{\Gamma(g_{ik})}u_{ik}^{g_{ik}-1}e^{-h_{ik}u_{ik}}\end{aligned} \quad (39)$$

and

$$\begin{aligned}f(\mathbf{V})&=\prod_{i=1}^{I}\prod_{k=1}^{K}\text{Gam}(v_{ik};s_{ik},t_{ik})\\&=\prod_{i=1}^{I}\prod_{k=1}^{K}\frac{t_{ik}^{s_{ik}}}{\Gamma(s_{ik})}v_{ik}^{s_{ik}-1}e^{-t_{ik}v_{ik}}.\end{aligned} \quad (40)$$

Therefore, the joint density of latent variables $\mathbf{\Theta}=\{\mathbf{Z},\mathbf{U},\mathbf{V}\}$ and observations $\mathbf{Y}$ given $\mathbf{\Pi}$ can be written as

$$\begin{aligned}&f(\mathbf{Y},\mathbf{\Theta}|\mathbf{\Pi})\\&=f(\mathbf{Y}|\mathbf{Z},\mathbf{U},\mathbf{V})f(\mathbf{Z}|\mathbf{\Pi})f(\mathbf{U})f(\mathbf{V})\\&=\prod_{n=1}^{N}\prod_{i=1}^{I}\left[\prod_{k=1}^{K}\frac{\Gamma(u_{ik}+v_{ik})}{\Gamma(u_{ik})\Gamma(v_{ik})}y_{nk}^{u_{ik}-1}(1+y_{nk})^{-(u_{ik}+v_{ik})}\right]^{z_{ni}}\\&\quad\times\prod_{n=1}^{N}\prod_{i=1}^{I}\pi_i^{z_{ni}}\\&\quad\times\prod_{i=1}^{I}\prod_{k=1}^{K}\left[\frac{h_{ik}^{g_{ik}}}{\Gamma(g_{ik})}u_{ik}^{g_{ik}-1}e^{-h_{ik}u_{ik}}\cdot\frac{t_{ik}^{s_{ik}}}{\Gamma(s_{ik})}v_{ik}^{s_{ik}-1}e^{-t_{ik}v_{ik}}\right]\end{aligned} \quad (41)$$

By applying the EVI method [4], [5], [29], we can acquire the analytically tractable solution to Bayesian estimation of a GiDMM, which is summarized in Algorithm 1. The related expectations in Algorithm 1 are calculated as

$$\begin{aligned}&\ln\rho_{ni}\\&=\ln\pi_i+\sum_{k=1}^{K}\Big[\tilde{\mathcal{R}}_i+(\bar{u}_{ik}-1)\ln y_{nk}-(\bar{u}_{ik}+\bar{v}_{ik})\ln(1+y_{nk})\Big],\end{aligned}$$

$$\begin{aligned}&\tilde{\mathcal{R}}_i\\&=\ln\frac{\Gamma(\bar{u}_{ik}+\bar{v}_{ik})}{\Gamma(\bar{u}_{ik})\Gamma(\bar{v}_{ik})}+[\psi(\bar{u}_{ik}+\bar{v}_{ik})-\psi(\bar{u}_{ik})]\left[\overline{\ln u_{ik}}-\ln\bar{u}_{ik}\right]\bar{u}_{ik}\\&\quad+[\psi(\bar{u}_{ik}+\bar{v}_{ik})-\psi(\bar{v}_{ik})]\left[\overline{\ln v_{ik}}-\ln\bar{v}_{ik}\right]\bar{v}_{ik},\end{aligned}$$

and

$$\begin{aligned}\bar{u}_{ik}&=\frac{g_{ik}^*}{h_{ik}^*},\quad\overline{\ln u_{ik}}=\psi(g_{ik}^*)-\ln h_{ik}^*,\\\bar{v}_{ik}&=\frac{s_{ik}^*}{t_{ik}^*},\quad\overline{\ln v_{ik}}=\psi(s_{ik}^*)-\ln t_{ik}^*.\end{aligned} \quad (42)$$

The point estimation of the GiDMM parameters can be obtained by taking the posterior means as

$$\hat{u}_{ik}=\frac{g_{ik}^*}{h_{ik}^*},\hat{v}_{ik}=\frac{s_{ik}^*}{t_{ik}^*},i=1,\cdots,I,k=1,\cdots,K. \quad (43)$$

In addition, the mixing coefficients are given by

$$\pi_i=\frac{1}{N}\sum_{n=1}^{N}\mathbf{E}[z_{ni}] \quad (44)$$

**Algorithm 1** Variational GiDMM.

**Input:** Observation $\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$, initial number of mixture components $I$
Initialize $g_{ik}=1$, $h_{ik}=0.1$, $s_{ik}=1$, $t_{ik}=0.1$ for $i=1,\cdots,I,k=1,\cdots,K$
Set $y_{nk}=x_{nk}/(1+\sum_{l=1}^{K-1}x_{nl})$
**repeat**
$\mathbf{E}[z_{ni}]=\rho_{ni}/\sum_{i=1}^{I}\rho_{ni}$
$g_{ik}^*=g_{ik_0}+\sum_{n=1}^{N}\mathbf{E}[z_{nk}]\,[\psi(\bar{u}_{ik}+\bar{v}_{ik})-\psi(\bar{u}_{ik})]\,\bar{u}_{ik}$
$h_{ik}^*=h_{ik_0}-\sum_{n=1}^{N}\mathbf{E}[z_{nk}]\,[\ln y_{nk}-\ln(1+y_{nk})]$
$s_{ik}^*=s_{ik_0}+\sum_{n=1}^{N}\mathbf{E}[z_{nk}]\,[\psi(\bar{u}_{ik}+\bar{v}_{ik})-\psi(\bar{v}_{ik})]\,\bar{v}_{ik}$
$t_{ik}^*=t_{ik_0}-\sum_{n=1}^{N}\mathbf{E}[z_{nk}]\ln(1+y_{nk})$
**until** Stop criteria are reached.
**Output:** The optimal hyper-parameters $g_{ik}^*$, $h_{ik}^*$, $s_{ik}^*$, $t_{ik}^*$.

## REFERENCES


[1] S. Park, E. Serpedin, and K. Qaraqe, "Gaussian assumption: The least favorable but the most useful," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 183–186, May 2013.

[2] Z. Ma, "Non-Gaussian statistical models and their applications," Ph.D. dissertation, KTH - Royal Institute of Technology, 2011.

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

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

[5] Z. Ma, A. 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. 876–889, 2015.

[6] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte carlo," in *Proceedings of International Conference on Machine Learning*, 2008, pp. 880–887.

[7] Z. Ma and A. E. Teschendorff, "A variational Bayes beta mixture model for feature selection in DNA methylation studies," *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 4, 2013.

[8] 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.

[9] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 998–1011, May 2013.

[10] K. Mammasis, R. W. Stewart, and J. S. Thompson, "Spatial fading correlation model using mixtures of von Mises Fisher distributions," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 2046–2055, April 2009.

[11] 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.

[12] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-Gaussian distributions," *IET Signal Processing*, vol. 6, no. 7, pp. 684–688, September 2012.

[13] D. Xu, C. Shen, and F. Shen, "A robust particle filtering algorithm with non-Gaussian measurement noise using student-t distribution," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 30–34, 2014.

[14] A. Amini, P. Thevenaz, J. Ward, and M. Unser, "On the linearity of Bayesian interpolators for non-Gaussian continuous-time AR(1) processes," *IEEE Transactions on Information Theory*, vol. 59, no. 8, pp. 5063–5074, Aug 2013.

[15] Z. Xu, S. MacEachern, and X. Xu, "Modeling non-Gaussian time series with nonparametric Bayesian model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 372–382, Feb 2015.

[16] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web - Internet and Web Information Systems*, 2018.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, Ltd, 2000.

[19] C. Gong, T. Liu, D. Tao, K. Fu, E. Tu, and J. Yang, "Deformed graph laplacian for semisupervised learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2261–2274, Oct 2015.

[20] C. Gong, D. Tao, K. Fu, and J. Yang, "Fick's law assisted propagation for semisupervised learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 2148–2162, Sept 2015.

[21] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3249–3260, July 2016.

[22] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang, "Label propagation via teaching-to-learn and learning-to-teach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1452–1465, June 2017.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[24] S. M. Stigler, "Thomas Bayes's Bayesian inference," *Journal of the Royal Statistical Society. Series A (General)*, vol. 145, no. 2, pp. 250–258, 1982.

[25] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," 2004, pp. 41–62.

[26] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[27] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[28] W. Fan, N. Bouguila, and D. Ziou, "Variational learning for finite Dirichlet mixture models and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 762–774, May 2012.

[29] 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, Sep. 2014.

[30] 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, Jan 2010.

[31] 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.

[32] M. A. Mashrgy, N. Bouguila, and K. Daoudi, "A statistical framework for positive data clustering with feature selection: Application to object detection," in *European Signal Processing Conference (EUSIPCO)*, Sept 2013, pp. 1–5.

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

[34] 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, Sept 2016.

[35] D. M. Blei, "Probabilistic models of text and images," Ph.D. dissertation, University of California, Berkeley, 2004.

[36] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[37] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proceedings of the International Conference on Machine Learning*, 2010.

[38] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *Journal of the American Statistical Association*, vol. 105, pp. 324–335, 2010.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[40] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience.*, vol. 2009, no. CUED/F-INFENG/TR.609, July 2009.

[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] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advances in Mean Field Methods.*, M. Opper and D. Saad, Eds. MIT Press., 2001, pp. 129–159.

[43] F. Huszár, "Variational inference using implicit distributions," *arXiv PrePrint*, 2017.

[44] 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.

[45] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, Nov 2010.

[46] C. Archambeau, B. Lakshminarayanan, and G. Bouchard, "Latent IBP compound Dirichlet allocation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 321–333, Feb 2015.

[47] F. Zhuang, P. Luo, Z. Shen, Q. He, Y. Xiong, Z. Shi, and H. Xiong, "Mining distinction and commonality across multiple domains using generative model for text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 2025–2039, Nov 2012.

[48] T. Bdiri, N. Bouguila, and D. Ziou, "Variational Bayesian inference for infinite generalized inverted Dirichlet mixtures with feature selection and its application to clustering," *Applied Intelligence*, vol. 44, no. 3, pp. 507–525, 2016.

[49] X. Quan, L. Wenyin, and B. Qiu, "Term weighting schemes for question categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 1009–1021, May 2011.

[50] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive bayes for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2508–2521, Sept 2016.

[51] Y. Li and A. K. Jain, "Classification of text documents," *The Computer Journal*, vol. 8, no. 41, pp. 537–546, 1998.

[52] N. Bouguila, "Infinite Liouville mixture models with application to text and texture categorization," *Pattern Recognition Letters*, vol. 33, pp. 103–110, 2012.

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

[54] N. Bouguila, "Count data modeling and classification using finite mixtures of distributions," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 186–198, Feb 2011.

[55] B. Tang, H. He, P. M. Baggenstoss, and S. Kay, "A Bayesian classification approach using class-specific features for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1602–1606, June 2016.

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

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

[58] 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*, 2016, under review.

[59] P. Tirdad, N. Bouguila, and D. Ziou, *Variational Learning of Finite Inverted Dirichlet Mixture Models and Applications*. Cham: Springer International Publishing, 2015, pp. 119–145.

[60] R. C. Luo and C. C. Lai, "Multi-sensor fusion-based concurrent environment mapping and moving object detection for intelligent service robotics," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 8, pp. 4043–4051, Aug 2014.

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

[62] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2531–2539.

[63] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, June 2015.

[64] 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.

[65] Y. Pang, K. Zhang, Y. Yuan, and K. Wang, "Distributed object detection with linear SVMs," *IEEE Transactions on Cybernetics*, vol. 44, no. 11, pp. 2122–2133, Nov 2014.

[66] R. Lan and Y. Zhou, "Quaternion-Michelson descriptor for color image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5281–5292, Nov 2016.

[67] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2071–2084, Oct 2015.

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

[69] 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.

[70] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, June 2003, pp. II–409–15 vol.2.

[71] J. A. Rodriguez-Serrano, D. Larlus, and Z. Dai, "Data-driven detection of prominent objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1969–1982, Oct 2016.

[72] W. Viana, R. Braga, F. D. A. Lemos, J. M. O. de Souza, R. A. F. Carmo, R. M. C. Andrade, and H. Martin, "Mobile photo recommendation and logbook generation using context-tagged images," *IEEE MultiMedia*, vol. 21, no. 1, pp. 24–34, Jan 2014.

[73] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 276–288, Jan 2017.

[74] R. G. Cinbis, J. Verbeek, and C. Schmid, "Approximate Fisher kernels of non-iid image models for image categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 6, pp. 1084–1098, June 2016.

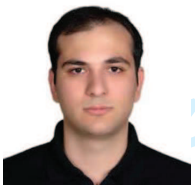
[75] S. Li, T. Lu, L. Fang, X. Jia, and J. A. Benediktsson, "Probabilistic fusion of pixel-level and superpixel-level hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7416–7430, Dec 2016.

[76] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.


**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.

**Jiyang Xie** is currently pursuing his Ph.D. degree with Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in image processing, data mining and deep learning.

**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.



**Jalil Taghia** has been a Postdoctoral Researcher at Cognitive & Systems Neuroscience at Stanford University School of Medicine since August 2015. He was a Postdoctoral Research Fellow in the Neural Information Processing Group at the Technical University of Berlin from October 2014 to July 2015. He received his Ph.D. in telecommunications from KTH (Royal Institute of Technology), Sweden, in 2014. His research interest lies in probabilistic inference in machine learning and statistics including approximate methods in Bayesian inference and directional statistics.

**Jing-Hao Xue** received the Dr. Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. Since 2008 he has worked in the Department of Statistical Science at University College London as a Lecturer and Senior Lecturer. His current research interests include statistical classification, high-dimensional data analysis, computer vision, and pattern recognition.

**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.