On the Convergence of Extended Variational Inference for Non-Gaussian Statistical Models

Zhanyu Ma, Jalil Taghia, and Jun Guo

Abstract—Variational inference (VI) is a widely used framework in Bayesian estimation. For most of the non-Gaussian statistical models, it is infeasible to find an analytically tractable solution to estimate the posterior distributions of the parameters. Recently, an improved framework, namely the extended variational inference (EVI), has been introduced and applied to derive analytically tractable solutions by employing lower-bound approximation to the variational objective function. Two conditions required for EVI implementation, namely the weak condition and the strong condition, are discussed and compared in this paper. In practical implementation, the convergence of the EVI depends on the selection of the lower-bound approximation, no matter with the weak condition or the strong condition. In general, two approximation strategies, the single lower-bound (SLB) approximation and the multiple lower-bounds (MLB) approximation, can be applied to carry out the lower-bound approximation. To clarify the differences between the SLB and the MLB, we will also discuss the convergence properties of the aforementioned two approximations. Extensive comparisons are made based on some existing EVI-based non-Gaussian statistical models. Theoretical analysis are conducted to demonstrate the differences between the weak and the strong conditions. Qualitative and quantitative experimental results are presented to show the advantages of the SLB approximation.

Index Terms—Bayesian estimation, non-Gaussian statistical models, variational inference, extended variational inference, lower-bound approximation, convergence

1 INTRODUCTION

Gaussian distribution is the ubiquitous probability distribution used in statistics, signal processing, and pattern recognition areas [1]. However, not all the data being processed are Gaussian distributed [2]. In many real-life applications, the distribution of data is asymmetric and, therefore, is not Gaussian distributed [3]. For example, the image pixel values [4], [5], the reviewer’s rating to an item in a recommendation system [6]–[8], and the DNA methylation level data [9] are distributed in a range with bounded support. The diversity gain over the $K_G$ fading [10] and the periodogram coefficients in speech enhancement [11], [12] are semi-bounded (nonnegative). The spatial fading correlation [13] and the yeast gene expressions [14] have directional property so that the $l_1$ norm equals one. In signal processing, the acoustic noise with colored spectra [15] and the measurement noise in state-space model [16] are heavy-tailed. In the stock market, the asymptotic behavior of the first-order autoregressive (AR) process is clearly non-Gaussian [17] and the underlying Bayesian copula model for the stock index series are non-Gaussian as well [18]. Although the above mentioned data represent diverse characteristics, a common property is that, these data not only have specific support range, but also have non-bell distribution shape. The natural properties of Gaussian distribution (the definition domain is unbounded and the distribution shape is symmetric) do not fit such data well. Hence, these data are non-Gaussian distributed. It has been found in recent studies that explicitly utilizing the non-Gaussian characteristics can significantly improve the practical performance [2], [4], [5], [9]–[16]. Hence, it is of particular importance and interest to make thorough studies of the non-Gaussian data and non-Gaussian statistical models.

Bayesian analysis plays an essential role in parameter estimation of statistical models [19]–[22]. Unlike the conventionally used maximum-likelihood (ML) estimation [23], Bayesian estimation assumes that the parameters potentially follow underlying 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 [21], [25]. Estimation of the posterior distribution via Bayesian estimation has several advantages over the ML estimation. Firstly, it gives statistical description to the parameters, rather than the simple point estimate that yield by the ML estimation. This makes Bayesian estimation more robust and reliable, by including the resulting uncertainty into the estimation [22]. Secondly, it can potentially prevent the overfitting problem, which is one of the drawbacks the ML estimation suffers. This is mainly due to the advantage of Occam’s razor effect in Bayesian estimation. Last but not the least, Bayesian estimation can estimate the model complexity automatically from the data. In ML estimation, model complexity decision usually requires cross validations and, therefore, is computationally costly [21], [23].

Variational inference (VI) framework, among others, is a widely used strategy to infer the posterior distribution of the parameters in Bayesian analysis [24], [26]–[28]. In a full Bayesian model where all the parameters are assigned with prior distributions, we minimize the Kullback-Leibler (KL) divergence of the true posterior distribution from the approximating one to obtain an optimal approximation to the posterior distribution [21, Ch. 10]. This procedure is equivalent to maximizing the lower-bound to the marginal likelihood (model evidence). The optimal posterior distribu-
tion can be obtained by iteratively updating one variable (or one variable group) while fixing the rest. However, unlike the famous Gaussian distribution \[11, 21, 29\], most of the non-Gaussian statistical models (e.g., beta mixture model (BMM) \[4, 6\], Dirichlet mixture model (DMM) \[31\], von-Mises Fisher mixture model (VMM) \[14\], beta-Gamma non-negative matrix factorization (BG-NMF) \[6\]) do not have analytically tractable solution to estimate the posterior distribution of the parameters. Numerical methods, e.g., Newton-Raphson algorithm, Gibbs sampling, Markov Chain Monte Carlo, are usually employed to sample from the posterior distribution \[5, 27\]. Numerical method often depends on Markov chain convergence and is in general computationally costly, especially in the high-dimensional space \[32\].

Recently, an improved framework, namely the extended variational inference (EVI) \[4, 6, 14, 31, 33\], has become popular in solving the above mentioned problem. Similar as the VI framework, EVI also seeks an optimal lower-bound approximation to the posterior distribution. The difference is that EVI relaxes the objective function (the evidence lower-bound) by constructing lower-bound approximation to the objective function. This lower-bound relaxation, which uses the convexity or relative convexity \[35\] of the objective function, can yield analytically tractable solution so that the parameter estimation is facilitated. Although 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 \[4, 6, 14, 33\]. In Bayesian estimation of BMM, Ma et al. \[4\] derived an analytically tractable solution which outperforms the numerical Gibbs sampling based method \[5\]. As an extension work, Bayesian estimation of DMM via EVI have been proposed in \[31\], 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 to provide lower-bound approximation \[14\]. For non-negative matrix factorization (NMF), EVI was also applied in deriving analytically tractable solutions for Poisson process (discrete) NMF \[38\], Gamma process NMF in music recording \[34\], and beta-Gamma NMF for bounded support data \[6\].

Convergence is an important issue in parameter estimation algorithm. For VI-based method, the objective function maximized during each iteration is convex or relatively convex in terms of the target variable’s posterior distribution \[21, 35\]. Hence, convergence is theoretically guaranteed. In EVI, the introduced lower-bound approximation to the objective function can be obtained either via 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 \[6, 14, 34\] and the other is the multiple lower-bounds (MLB) approximation \[4, 31\]. For EVI with SLB approximation, convergence is also guaranteed because the original objective function is replaced by one single lower-bound and this new objective function (i.e., the single lower-bound to the original objective function in VI) is convex (or relatively convex) and maximized during each iteration. However, when applying EVI with MLB approximation, the variable group is divided into different disjoint subsets and there exists 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 can not be theoretically guaranteed.

In order to clarify the convergence property of the EVI framework, we will discuss and summarize the conditions that required in EVI implementation. The SLB and MLB approximations will also be analyzed and compared qualitatively and quantitatively. Experimental results based on the recently proposed EVI-based BMM and DMM estimation algorithms will be presented to demonstrate the advantages of the SLB approximations. We draw some conclusions in the end.

## 2 Variational Inference and Extended Variational Inference

### 2.1 Variational Inference

In Bayesian estimation, a universal solution to the variational inference (VI) framework \[26\] is to approximate the posterior distribution by a product of several factor distributions and then update each factor distribution individually \[21\]. This method is the so-called factorized approximation (FA) which was developed from the mean field theory in physics \[40\]. With the FA method, the variational objective function that we want to maximize can be represented as the negative KL divergence as

\[
\mathcal{L} = \mathbb{E}_Z [\ln p(X, Z) - \ln q(Z)],
\]

where \(X\) is the observed data, \(Z\) denotes all the random variables. If \(Z\) can be (approximately) factorized into \(M\) disjoint groups as \(Z = \{Z_1, \ldots, Z_i, \ldots, Z_M\}\) and we approximate the true posterior distribution \(p(Z|X)\) as

\[
p(Z|X) \approx q(Z) = \prod_{i=1}^{M} q_i(Z_i),
\]

the optimal solution can be written as

\[
\ln q_i^*(Z_i) = \mathbb{E}_{Z \setminus Z_i} [\ln p(X, Z)] + \text{const.}
\]

The operator \(\mathbb{E}_{Z \setminus Z_i}\) means expectation with respect all the variables in \(Z\), except for \(Z_i\). If the optimal solution to the posterior distribution of \(Z_i\), which is \(\ln q_i^*(Z_i)\) in (3), has the same logarithmical form as the prior distribution, the conjugate match between the prior and the 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 \[4, 6, 31\]. This is due to the fact that the optimal solution depends on the expectation computed with respect to the factor distribution \(Z_i\).

### 2.2 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 moments \[39\] to derive an alternative for the objective function to simplify the calculation. Blei et al. \[33\] 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 idea was also applied in [41, 14, 21] for approximating the posterior distributions in BMM, DMM, and VMM, respectively. Using Jensen’s inequality has become commonplace in variational inference. In [34], 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 approximately calculated the posterior distribution. Moreover, the EVI strategy was also applied in low rank matrix approximation area [6], where the Taylor expansion and Jensen’s inequality were both applied for the purpose of deriving analytically tractable solution. All the aforementioned works utilized the following property. Given an auxiliary function which satisfies

\[
E_Z \left[ \ln p(X, Z) \right] \geq E_Z \left[ \ln \tilde{p}(X, Z) \right],
\]

the variational objective function (see [21], pp. 465 for more details) can be lower-bounded as

\[
\mathcal{L} = E_Z \left[ \ln p(X, Z) \right] - E_Z \left[ \ln q(Z) \right] \\
\geq E_Z \left[ \ln \tilde{p}(X, Z) \right] - E_Z \left[ \ln q(Z) \right] \\
\triangleq \tilde{\mathcal{L}}.
\]

Then we can maximize \(\tilde{\mathcal{L}}\), which is an lower-bound to the original objective function \(\mathcal{L}\) to asymptotically reach the maximum value of \(\mathcal{L}\). The approximated optimal solution in this case is written as

\[
\ln \tilde{q}(Z_i) = E_{Z \backslash Z_i} \left[ \ln \tilde{p}(X, Z) \right] + \text{const}.
\]

This method is the so-called EVI framework [4], [6], [14], [31], [33]–[35]. 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.

3 CONVERGENCE OF EVI

3.1 Weak Condition and Strong Condition

As mentioned in Sec. 2.2 finding an auxiliary function \(\tilde{p}(X, Z)\) 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

\[
p(X, Z) \geq \tilde{p}(X, Z).
\]

It is obvious that an auxiliary function satisfies (7) should also satisfy (4). Hence, the condition in (4) is named as the weak condition and the one in (7) 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 incurred by the applying weak condition is relatively smaller than that introduced by using the strong condition. Fig. 1 illustrates the different gaps introduced by the weak and strong conditions, respectively.

It is worthwhile to note that the auxiliary function \(\tilde{p}(X, Z)\) is not necessary to be a normalized probability density function (PDF). This will not affect the final solution

1. Actually, an auxiliary function that satisfies the strong condition cannot be a normalized PDF, as \(p(X, Z)\) itself is a normalized PDF.

2. According to our experience, globally concavity holds only for Gaussian distribution. For (most of) the non-Gaussian statistical models, the original function is not globally concave.
functions. For example, we could have \( M \) auxiliary functions as

\[
p(X, Z) \geq \tilde{p}_1(X, Z_1) \\
\vdots \\
p(X, Z) \geq \tilde{p}_i(X, Z_i) \\
\vdots \\
p(X, Z) \geq \tilde{p}_M(X, Z_M).
\]

(8)

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. 3.1, the optimal solution in (6) is

\[
\ln \tilde{q}_i^* (Z_i) = E_{X, Z_i}[\ln \tilde{p}_i(X, Z_i)] + \text{const.} \tag{9}
\]

In this case, the new objective function that maximized during each iteration is not unique. Hence, there is no globally objective function that is maximized during each iteration. The convergence cannot be theoretically guaranteed. Such procedure has been applied in [4] and [31]. Although it is not guaranteed theoretically, the convergence was observed empirically.

Let’s study a simple case with two disjoint groups in the MLB approximation. Assuming that \( Z = \{Z_1, Z_2\} \) and we have two auxiliary functions \( \tilde{p}_1(X, Z_1) \) and \( \tilde{p}_2(X, Z_2) \) for \( Z_1 \) and \( Z_2 \), respectively. As mentioned above, two different lower-bounds are obtained as

\[
\tilde{L}_1 = E_Z[\ln \tilde{p}_1(X, Z_1) - \ln q(Z)] \\
\tilde{L}_2 = E_Z[\ln \tilde{p}_2(X, Z_2) - \ln q(Z)]. \tag{10}
\]

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

\[
\ln \tilde{q}_1^* (Z_1) = E_{Z_1}[\ln \tilde{p}_1(X, Z_1)] + \text{const.} \tag{11a}
\]

\[
\ln \tilde{q}_2^* (Z_2) = E_{Z_2}[\ln \tilde{p}_2(X, Z_2)] + \text{const.} \tag{11b}
\]

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

\[
2 \times L \geq \tilde{L}_1 + \tilde{L}_2 \\
= E_Z[\ln \tilde{p}_1(X, Z_1)] - E_Z[\ln q(Z)] \tag{12a}
\]

\[
+ E_Z[\ln \tilde{p}_2(X, Z_2)] - E_Z[\ln q(Z)]. \tag{12b}
\]

When performing the update strategy \(11a\), we get \(12b\) to be maximized. This maximization makes the distribution of \( Z_1 \) to be less uncertain. As \( -E_Z[\ln q(Z)] \) in \(12c\) is the differential entropy of \( Z \), \(12c\) is decreasing while \(12b\) is maximizing. It is hard to evaluate if \(12b\) changes more than \(12c\) or not. Thus, the overall lower-bound, i.e., \( \tilde{L}_1 + \tilde{L}_2 \) in \(12a\), might decrease during some iterations. On the one hand, as the lower-bound (i.e., \( \tilde{L}_1 + \tilde{L}_2 \)) to the original objective function can not be guaranteed to be maximized all the time, this strategy may not promise convergence. On the other hand, if the change to \(12b\) is larger than to \(12c\), the convergency 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 underdetermined.

In summary, SLB approximation can theoretically guarantee the convergence while MLB approximation, in general, cannot promise convergence.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

Recently, several EVI-based parameter estimation algorithms for non-Gaussian statistical models have been proposed. Among others, the EVI-based Bayesian BMM [4] and the EVI-based Bayesian DMM [31] took the strong condition and the related analytically tractable solutions were derived with MLB approximation. With SLB approximation, the improved work about EVI-based Dirichlet mixture model was proposed. Regarding the non-negative matrix factorization, Hoffman et al. and Ma et al. proposed the EVI-based strategies for musical signal [34] and bounded support data [3], respectively. The EVI-based von Mises-Fisher mixture model was proposed in [13] where a structural factorization was considered. For all the aforementioned SLB approximation-based method, the weak condition is fulfilled.

In this section, 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.

4.1 Comparisons of Weak and Strong Condition

Since Dirichlet distribution is a multivariate case of beta distribution, the EVI-based Bayesian BMM that constructs

3. In practice (e.g., [4], [31]), the EVI-based algorithm may also converge with MLB approximation. However, it is empirical result without proof.
a auxiliary function with weak condition can be obtained based on the work in [37] by simply setting the dimension $K = 2$. The EVI-based Bayesian BMM proposed in [4] utilized the strong condition to choose the auxiliary function. We compare these two different methods to demonstrate the differences between the strong and weak conditions.

Following the same notation in [4], we denote a multivariate Bayesian BMM with observation data $X$ as

$$f(X; \Pi, U, V) = \prod_{i=1}^{N} \prod_{j=1}^{I} \pi_i \text{Beta}(x_{ni}; u_i, v_i), \quad (13)$$

where $\pi_i$ is the mixture weight for the $i$th mixture component and Beta($x; u, v$) is the beta distribution, which can be denoted as

$$\text{Beta}(x; u, v) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} x^{u-1}(1-x)^{v-1}, \quad u, v > 0. \quad (14)$$

We consider the observation $x_n$ and the unobserved indication vector $z_n$ as the complete data. The conditional distribution of $X = \{x_1, \ldots, x_N\}$ and $Z = \{z_1, \ldots, z_N\}$ given the latent variables $\{U, V, \Pi\}$ is

$$f(X, Z|U, V, \Pi) = f(X|U, V, \Pi, Z) f(Z|\Pi) = f(X|U, V, Z) f(Z|\Pi) = \prod_{n=1}^{N} \prod_{i=1}^{I} \{\pi_i \text{Beta}(x_{ni}|u_i, v_i)\}^{z_{ni}}. \quad (15)$$

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 for the posterior distributions, the most challenging part with the EVI framework is to calculate the expectation of the bivariate log-inverse-beta (LIB) function

$$E_{u_i, v_i}[\text{LIB}(u_i, v_i)] = E_{u_i, v_i} \left[ \frac{\Gamma(u_i + v_i)}{\Gamma(u_i)\Gamma(v_i)} \right]. \quad (16)$$

### 4.1.1 EVI-based Bayesian BMM with Weak Condition [37]

In the Bayesian BMM with SLB approximation [4] the new objective function that we are maximizing is

$$\mathbb{E}_Z [\ln \tilde{p}_w(X, Z)] = \bar{L}_{\text{SLB}}$$

$$= \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln u_1] - \ln \pi_1$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln v_1] - \ln \pi_1, \quad (17)$$

where $\bar{\pi}$ 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 $\mathbb{E}_Z [\ln p(X, Z)] \geq \mathbb{E}_Z [\ln \tilde{p}_w(X, Z)]$. Moreover, this lower-bound is identical for all the variables $u_i, v_i$, and $z_{ni}$.

4. A Bayesian BMM with SLB approximation can be derived from the Bayesian DMM with SLB approximation [37] by setting the dimensions of the Dirichlet variable equal to two.

### 4.1.2 EVI-based Bayesian BMM with Strong Condition [4]

For the case with strong condition, an auxiliary function $\tilde{p}_s(X, Z)$ 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

$$\tilde{p}_{s_u_i}(X, Z) = \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1)$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1) - \ln \pi_1, \quad (18)$$

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

$$\bar{L}_{\text{MLB}_{s_u_i}} = \mathbb{E}_Z [\tilde{p}_{s_u_i}(X, Z)]$$

$$= \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln u_i] - \ln \pi_1$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln v_i] - \ln \pi_1$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln u_i] - \ln \pi_1. \quad (19)$$

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

$$\bar{L}_{\text{MLB}_{s_v_i}} = \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln u_i] - \ln \pi_1$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \text{E} [\ln v_i] - \ln \pi_1$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln v_i] - \ln \pi_1. \quad (20)$$

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

$$\tilde{p}_{s_{z_{ni}}}(X, Z)$$

$$= \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1)$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1) - \ln \pi_1$$

$$+ 0.5 \cdot \bar{\pi}_1^2 [\psi'(\pi_1 + \pi_1) - \psi'(\pi_1)] \text{E} [\ln u_i] - \ln \pi_1]^2$$

$$+ 0.5 \cdot \bar{\pi}_1^2 [\psi'(\pi_1 + \pi_1) - \psi'(\pi_1)] \text{E} [\ln v_i] - \ln \pi_1]^2$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln u_i] - \ln \pi_1]$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln v_i] - \ln \pi_1]. \quad (21)$$

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

$$\bar{L}_{\text{MLB}_{s_{z_{ni}}}} = \mathbb{E}_Z [\tilde{p}_{s_{z_{ni}}}(X, Z)]$$

$$= \ln \frac{\Gamma(\pi_1 + \pi_1)}{\Gamma(\pi_1)\Gamma(\pi_1)}$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1)$$

$$+ \bar{\Gamma}_1 [\psi(\pi_1 + \pi_1) - \psi(\pi_1)] \ln(\pi_1) - \ln \pi_1$$

$$+ 0.5 \cdot \bar{\pi}_1^2 [\psi'(\pi_1 + \pi_1) - \psi'(\pi_1)] \text{E} [\ln u_i] - \ln \pi_1]^2$$

$$+ 0.5 \cdot \bar{\pi}_1^2 [\psi'(\pi_1 + \pi_1) - \psi'(\pi_1)] \text{E} [\ln v_i] - \ln \pi_1]^2$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln u_i] - \ln \pi_1]$$

$$+ \bar{\Gamma}_1 \cdot \psi'(\pi_1 + \pi_1) \text{E} [\ln v_i] - \ln \pi_1]. \quad (22)$$

It has been analyzed in Sec. 3.1 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 strong condition is also a MLB approximation. We focus only on the comparisons of weak and strong conditions in this section. The comparisons about 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

\[
\Delta \hat{L}_{\text{SLB vs. MLB}} = \hat{L}_{\text{SLB}} - \hat{L}_{\text{MLB}},
\]

where we used the fact that \( \psi'(x) > 0 \) and \( \ln x \) is a convex function in terms of \( x \). For \( v_i \), it is straightforward to show the difference is also positive by using the symmetric properties.

When comparing \( \hat{L}_{\text{SLB}} \) with \( \hat{L}_{\text{MLB}} \), the difference is

\[
\Delta \hat{L}_{\text{SLB vs. MLB}} = \hat{L}_{\text{SLB}} - \hat{L}_{\text{MLB}},
\]

\[
= -\left\{0.5 \cdot \frac{\partial}{\partial u_i} \psi(\bar{u}_i + \bar{v}_i) - \psi'(\bar{u}_i)\right\} \mathbb{E} \left[ \ln u_i - \ln \bar{u}_i \right]^2 + 0.5 \cdot \frac{\partial}{\partial \pi_i} \left[ \psi(\bar{u}_i + \bar{v}_i) - \psi'(\bar{pi}_i)\right] \mathbb{E} \left[ \ln v_i - \ln \bar{u}_i \right]^2 + \bar{v}_i \cdot \frac{\partial}{\partial \pi_i} \psi(\bar{u}_i + \bar{v}_i) \mathbb{E} \left[ \ln u_i - \ln \bar{v}_i \right] - \mathbb{E} \left[ \ln v_i - \ln \bar{v}_i \right].
\]

It can be proved that the difference \( \Delta \hat{L}_{\text{SLB vs. MLB}} \) is also greater than or equal to 0. More details for this proof can be found in Appendix ??.

The aforementioned three positive differences indicate that the new objective function with weak condition is tighter (i.e., closer to the original objective function) than that with strong condition. 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.

Similar analysis can be applied to the Bayesian DMM with MLB and the Bayesian DMM with SLB, as Dirichlet distribution is an multivariate extension of beta distribution.

### 4.2 Comparisons of MLB and SLB Approximations

In the previous section, we analyzed and compared the weak and the strong conditions for the EVI framework. Another important issue in EVI implementation is to distinguish the MLB and the SLB approximations, as the latter one cannot guarantee convergence but the first one may not. To this end, we compare the MLB approximation-based algorithm with the SLB approximation-based algorithm in this section.

#### 4.2.1 Observations of Non-convergence

As discussed in Sec. ??, the convergence of the MLB method is not guaranteed. We ran the MLB approximation-based Bayesian BMM algorithm and Bayesian DMM algorithm, respectively, and monitored the value of the objective function during each iteration. It can be observed that, for some rounds of simulations, the objective function is decreasing during some iterations. Although this non-convergence can be observed in some of the simulation rounds (2 ~ 3 times out of 10 rounds of simulations), in practice, this fact indicates that the MLB approximation-based method may not promise convergence. Figure ?? lists the decreasing objective function values and the corresponding iterations. For the SLB approximation-based Bayesian BMM and Bayesian DMM, the monitored objective function was always increasing until converging. The observation of non-convergence demonstrates that the convergence with MLB approximation is undetermined.

#### 4.2.2 Comparisons of Estimation Accuracy

In this section, we compare the MLB approximation with the SLB approximation quantitatively. With a known BMM or DMM, 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 to examine which approximation is better. With the obtained posterior distribution \( q^*(Z) \), the original variational objective function is calculated numerically by sampling method. Hence, we got two different values, \( \Delta L_{\text{SLB}} \) and \( \Delta L_{\text{MLB}} \), from the MLB approximation and the SLB 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 \( \text{KL}(p(X|\Theta)||\hat{p}(X|\hat{\Theta})) \), where \( \Theta \) is the true parameter vector and \( \hat{\Theta} \) is the estimated one. Similarly, we numerically calculated KL with and KL from the SLB and MLB approximations. The smaller the KL divergence is, the more accurate the estimation is.

For Bayesian BMM, the comparisons are presented in Table ?? and Figure ?? The simulations were run 20 rounds and the mean values are reported. The comparisons of the Bayesian DMM via SLB and MLB approximations are illustrated in Table ?? and Figure ??.

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

6. For the MLB approximation, we only take those simulation rounds that always converge into consideration.
that the SLB approximation can theoretically guarantee
model and Bayesian Dirichlet mixture model demonstrated
quantitative evaluations based on Bayesian beta mixture
showed that the weak condition, in general, can incur
lower-bounds (MLB) approximation. Theoretical analysis
framework. Moreover, we also analyzed and compared the
SLB approximation is superior to the MLB approximation.

Table 2
Comparisons of the objective functions for Bayesian BMM.

| Model | Parameters | SLB - MLB | KL SLB - MLB |
|-------|------------|-----------|--------------|
| A     | π₁ = 0.3, v₁ = 2, v₂ = 8 | 3.6 × 10⁻³ | -2.8 × 10⁻³ |
|       | π₂ = 0.7, v₁ = 10, v₂ = 4 |          |              |
| B     | π₁ = 0.6, v₁ = 10, v₂ = 2 | 1.3 × 10⁻³ | -0.58 × 10⁻³ |
|       | π₂ = 0.4, v₁ = 2, v₂ = 12 |          |              |
|       | π₃ = 0.3, v₁ = 10, v₂ = 10 |          |              |

Fig. 4. Comparisons of the original objective functions in Bayesian BMM. The central mark is the median, the edges are the 25th and 75th
percentiles. The outliers are marked individually. Model settings are the
same as Table 2.

approximation yields higher objective function value than the
MLB approximation. Meanwhile, the KL divergences
obtained by the SLB approximation are all smaller than
those obtained by the MLB. These facts demonstrate that
the SLB approximation is superior to the MLB approximation.

5 CONCLUSIONS
The extended variational inference (EVI) framework can be
applied in efficiently 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
single lower-bound (SLB) approximation and the multiple
lower-bounds (MLB) approximation. Theoretical analysis
showed that the weak condition, in general, can incur
smaller systematic gap than the strong condition. Hence, the
weak condition is more preferable in practice. Furthermore,
quantitative evaluations based on Bayesian beta mixture
model and Bayesian Dirichlet mixture model demonstrated
that the SLB approximation can theoretically guarantee
convergence and is superior to the MLB approximation.

REFERENCES
[1] S. Park, E. Serpedin, and K. Qaraqe, “Gaussian assumption: The
least favorable but the most useful,” IEEE Signal Processing Maga-
zine, 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 Cy-
bernetics, 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] N. Bouguela, D. Ziou, and E. Monga, “Practical Bayesian esti-
ination of a finite beta mixture through Gibbs sampling and its
applications,” Statistics and Computing, vol. 16, pp. 215–225, 2006.
[6] 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.
[7] 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.
[8] ——, “Probabilistic matrix factorization,” in Proceedings of Ad-
vanced Neural Information Processing Systems, 2008, pp. 1257–1264.
[9] Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes, “Application of
beta mixture models in bioinformatics,” Bioinformatics applications
note, vol. 21, pp. 2118–2122, 2005.
[10] 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.
[11] N. Mohammadiha, R. Martin, and A. Leijon, “Spectral domain
speech enhancement using HMM state-dependent super-Gaussian
priors,” IEEE Signal Processing Letters, vol. 20, no. 3, pp. 253–256,
March 2013.
[12] N. Mohammadiha and A. Leijon, “Nonnegative HMM for babble
noise derived from speech HMM: Application to speech enhance-
ment,” IEEE Transactions on Audio, Speech, and Language Processing,
vol. 21, no. 5, pp. 998–1011, May 2013.
[13] 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,
p. 2046–2055, April 2009.
[14] 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.
[15] L. Zhao 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.
[16] 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.
[17] 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,
p. 5063–5074, Aug 2013.
[18] 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.
[19] K. Fukunaga, Introduction to Statistical Pattern Recognition.
Academic Press, 1990.
[20] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recog-
nition: a review,” IEEE Transactions on Pattern Analysis and Machine
Intelligence, vol. 22, pp. 4–37, 2000.
[21] C. M. Bishop, Pattern Recognition and Machine Learning. Spring-
er, 2006.
[22] J. M. Bernardo and A. F. M. Smith, Bayesian Theory. John Wiley &
Sons, Ltd, 2000.
[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likely-
hood 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. F. Tipping, “Bayesian inference: An introduction to principles
and practice in machine learning,” 2004, pp. 41–62.
[26] 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.
[27] D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet
process mixtures,” Bayesian Analysis, vol. 1, pp. 121–144, 2006.
[28] C. W. Fox and S. J. Roberts, “A tutorial on variational Bay es
inference,” Machine Learning, vol. 22, pp. 4–37, 2000.
[29] J. T. Ormerod and M. P. Wand, “Gaussian variational approximate
inference for generalized linear mixed models,” Journal of Computa-
tional and Graphical Statistics, vol. 21, no. 1, pp. 2–17, 2012.
[30] J. Täghia and A. Leijon, “Variational inference for Watson mixture model,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 1886–1900, 2015.

[31] 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.

[32] D. M. Blei, “Probabilistic models of text and images,” Ph.D. dissertation, University of California, Berkeley, 2004.

[33] D. M. Blei and J. D. Lafferty, “Correlated topic models,” in Advances in Neural Information Processing Systems (NIPS), 2006.

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

[35] 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.

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

[37] Z. Ma, P. K. Rana, J. Täghia, 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.

[38] 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.

[39] P. J. Bickel and K. A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics. Pearson Prentice Hall, 2007.

[40] 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.