Pattern Recognition 47 (2014) 3143–3157

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Bayesian estimation of Dirichlet mixture model with variational inference

Zhanyu Ma [a,*], Pravin Kumar Rana [b], Jalil Taghia [b], Markus Flierl [b], Arne Leijon [b]

[a] Pattern Recognition and Intelligent Systems Lab., Beijing University of Posts and Telecommunications, Beijing, 100876, China
[b] KTH - Royal Institute of Technology, School of Electrical and Engineering, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form
1 February 2014
Accepted 1 April 2014
Available online 12 April 2014

Keywords:
Bayesian estimation
Variational inference
Extended factorized approximation
Relative convexity
Dirichlet distribution
Gamma prior
Mixture modeling
LSF quantization
Multiview depth image enhancement

ABSTRACT

In statistical modeling, parameter estimation is an essential and challengeable task. Estimation of the parameters in the Dirichlet mixture model (DMM) is analytically intractable, due to the integral expressions of the gamma function and its corresponding derivatives. We introduce a Bayesian estimation strategy to estimate the posterior distribution of the parameters in DMM. By assuming the gamma distribution as the prior to each parameter, we approximate both the prior and the posterior distribution of the parameters with a product of several mutually independent gamma distributions. The extended factorized approximation method is applied to introduce a single lower-bound to the variational objective function and an analytically tractable estimation solution is derived. Moreover, there is only one function that is maximized during iterations and, therefore, the convergence of the proposed algorithm is theoretically guaranteed. With synthesized data, the proposed method shows the advantages over the EM-based method and the previously proposed Bayesian estimation method. With two important multimedia signal processing applications, the good performance of the proposed Bayesian estimation method is demonstrated.

1. Introduction

Statistical modeling plays an important role in various research areas [1–3]. It provides a way to connect the data with the statistics. An essential part in statistical modeling is to estimate the values of the parameters in the distribution or to estimate the distribution of the parameters, if we consider them as random variables. The maximum likelihood (ML) estimation method gives point estimates to the parameters and disregards the remaining uncertainty in the estimation. Rather than taking the point estimates, the Bayesian estimation method gives the posterior probability distributions over all model parameters, using the observed data together with the prior distributions [3]. In general, compared to the ML estimation, the Bayesian estimation of the parameters in a statistical model could yield a robust and stable estimate, by including the resulting uncertainty into the estimation, especially when the amount of the observed data is small [4].

The Gaussian distribution and the corresponding Gaussian mixture model (GMM) are widely used to model the underlying distribution of the data. However, not all data we would like to model can be safely assumed to be Gaussian distributed [5]. Recently, the studies of non-Gaussian statistical models have become popular for the purpose of modeling bounded or semi-bounded data (see *e.g.*, [6–9]). The non-Gaussian statistical models include, among others, the beta distribution, the gamma distribution, and the Dirichlet distribution.

The Dirichlet distribution and the corresponding Dirichlet mixture model (DMM) were frequently applied to model proportional data, for example, in image processing [10], in text analysis [11], and in data mining [12]. For speech processing, applications of Dirichlet distribution in the line spectral frequency (LSF) parameter quantization [13] were shown superior to conventional GMM based methods. Another usage of the Dirichlet distribution is to model the probabilities of the weighting factors in a mixture model [14,15]. In non-parametric Bayesian modeling, the Dirichlet process is actually an infinite-dimensional generalization of the Dirichlet distribution so that an infinite mixture model can be obtained [15–17]. Here, we study only the finite DMM and the work conducted can also be extended to the infinite mixture modeling case.

In this paper, we carry on our previous study of Bayesian analysis of BMM [8] and extend it to the Bayesian analysis of DMM. The parameters in a Dirichlet distribution are assumed mutually independent and each of them is assigned by a gamma prior.

* Corresponding author.
E-mail address: mazhanyu@bupt.edu.cn (Z. Ma).

http://dx.doi.org/10.1016/j.patcog.2014.04.002
0031-3203/

Although this assumption violates the correlation among the parameters, it captures the non-negative properties of those parameters. By this assumption, we can apply the factorized approximation (FA) method to carry out the Bayesian estimation. However, as the expectation of the multivariate log-inverse-beta (MLIB) function cannot be calculated explicitly, an analytically tractable solution to the posterior distribution is not feasible. To overcome this problem, we study some relative convex properties of the MLIB function. Using these convexities, we approximate the expectation of the MLIB function by a single lower-bound (SLB). With this derived SLB and by principles of the VI framework and the extended factorized approximation (EFA) method [8,18–24], we approximate the posterior distributions of the parameters in a Dirichlet distribution with a product of several mutually independent gamma distributions, which satisfies the conjugate match between the prior and posterior distributions. Finally, an analytically tractable solution for calculating the posterior distribution is obtained. This analytically tractable solution avoids the numerical calculations in the EM algorithm [25,10].

The proposed method, which is a full Bayesian framework, can automatically determine the model complexity (in terms of the number of necessary mixture components) based on the data. This task is also challenging in model estimation and the ML estimation itself cannot handle this issue. Moreover, the overfitting problem in the ML estimation can also be prevented due to the advantages of Occam's razor effect in Bayesian estimation. With synthesized data evaluation, the effectiveness and the accuracy of the proposed Bayesian estimation method over the ML estimation method [10,25] and the recently proposed Bayesian estimation method [12] are demonstrated. For the real life applications, we evaluate the proposed Bayesian estimation method with two important multimedia signal processing applications, namely (1) the LSF parameter quantization in speech coding [13] and (2) the multi-view depth image enhancement in free-viewpoint television (FTV) [26]. For both applications, the proposed Bayesian method works well and shows improvement over the conventional methods.

The remaining parts of this paper are organized as follows: the DMM and the Bayesian analysis of a DMM are introduced in Sections 2 and 3, respectively. In Section 4, we show the efficiency and good performance of the proposed method with the synthesized data and the real life data. Some conclusions are drawn in Section 5.

2. Dirichlet mixture model

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[1]

$$\mathrm{Dir}(\mathbf{x};\mathbf{u})=\frac{\Gamma(\sum_{k=1}^{K+1}u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)}\prod_{k=1}^{K+1}x_k^{u_k-1},\quad u_k>0,\ 0<x_k<1, \tag{1}$$

where $x_{K+1}=1-\sum_{k=1}^{K}x_k$, $\mathbf{u}=[u_1,\ldots,u_{K+1}]^{\mathrm{T}}$ is the parameter vector, and $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}\,dt$. The shape of the Dirichlet distribution depends on the parameters. When $u_k>1$, $k=1,\ldots,K+1$, it is unimodally distributed. This is a typical case in practical applications. Thus in this paper, we study only the Dirichlet distribution with all its parameters greater than one.

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

To model the multimodality of the data, the mixture modeling technique [14] can be applied to create a DMM. With I mixture components, the PDF of a DMM can be represented, given a set of N *i.i.d.* observations $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]$, as

$$f(\mathbf{X};\boldsymbol{\Pi},\mathbf{U})=\prod_{n=1}^{N}\sum_{i=1}^{I}\pi_i\,\mathrm{Dir}(\mathbf{x}_n;\mathbf{u}_i),\quad \pi_i>0,\quad \sum_{i=1}^{I}\pi_i=1, \tag{2}$$

where $\boldsymbol{\Pi}=[\pi_1,\ldots,\pi_I]^{\mathrm{T}}$ is the mixture weights and $\mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_I]$ is the parameter matrix.

3. Bayesian estimation with variational inference framework

For a distribution belonged to the exponential family, the conjugate prior and the corresponding posterior distribution always exist [3]. Similar to the beta distribution [8], the Dirichlet distribution has its conjugate prior and the corresponding posterior distributions. However, they are not tractable in practical use. Thus we follow the principle of VI framework [18,3] to approximate the prior and posterior distributions. With the proposed approximation, the obtained prior and posterior distributions can be easily calculated and used.

3.1. Conjugate prior to Dirichlet distribution

Since the Dirichlet distribution is a member of the exponential family, the conjugate prior of the Dirichlet distribution exists. If we assume that the parameter vector $\mathbf{u}=[u_1,\ldots,u_{K+1}]^{\mathrm{T}}$ is a vector random variable, then the prior distribution of $\mathbf{u}$ can be expressed as

$$f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0)=\frac{1}{C(\boldsymbol{\beta}_0,\nu_0)}\left[\frac{\Gamma(\sum_{k=1}^{K+1}u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)}\right]^{\nu_0}e^{-\boldsymbol{\beta}_0^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_{K+1})}, \tag{3}$$

where $\boldsymbol{\beta}_0=[\beta_{1_0},\ldots,\beta_{K+1_0}]^{\mathrm{T}}$ and ν_0 are the hyperparameters in the prior distribution. $C(\boldsymbol{\beta}_0,\nu_0)$ is the normalization factor. $\mathbf{1}_m$ denotes an m-dimensional vector with all elements equal to one. With Bayes' theorem and combining (1) and (3) together, we can obtain the posterior distribution of the parameters, given the observation $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]$, as

$$\begin{aligned}f(\mathbf{u}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N)&=\frac{\mathrm{Dir}(\mathbf{X}|\mathbf{u})f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0)}{\int\mathrm{Dir}(\mathbf{X}|\mathbf{u})f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0)\,d\mathbf{u}}\\&=\frac{1}{C(\boldsymbol{\beta}_N,\nu_N)}\left[\frac{\Gamma(\sum_{k=1}^{K+1}u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)}\right]^{\nu_N}e^{-\boldsymbol{\beta}_N^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_{K+1})},\end{aligned} \tag{4}$$

where $\boldsymbol{\beta}_N=\boldsymbol{\beta}_0-\ln\mathbf{X}\times\mathbf{1}_N$, $\nu_N=\nu_0+N$ are the hyperparameters in the posterior distribution. Since some statistics of $\mathbf{u}$, *e.g.*, the mean, the covariance, cannot be obtained directly (by an analytically tractable expression) from (3) or (4), it is not convenient to use them in practical problems. In the following paragraphs, we will apply the VI framework to approximate the prior and posterior distributions of the parameters in a DMM. These approximations can lead to an analytically tractable solution and would be easily used in practice.

3.2. Factorized approximation

In a DMM, the observations $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]$ are considered as the incomplete data and an I-dimensional indication vector $\mathbf{z}_n=[z_{n1},\ldots,z_{nI}]^{\mathrm{T}}$ is assigned to each observation $\mathbf{x}_n$ to build a complete data set. Only one element in the indication vector is equal to 1 and the remaining elements are zeros. Thus $z_{ni}=1$ indicates that the nth observation is generated from the ith mixture component. For N observations, we have N indication vectors denoted as $\mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_N]$. If we treat all the parameters in (2) as the random variables, the conditional PDF of the complete

data $\{\mathbf{X},\ \mathbf{Z}\}$ given $\{\mathbf{\Pi},\ \mathbf{U}\}$ can be factorized as

$$f(\mathbf{X},\mathbf{Z}|\mathbf{\Pi},\mathbf{U}) = f(\mathbf{X}|\mathbf{Z},\mathbf{U})f(\mathbf{Z}|\mathbf{\Pi}), \tag{5}$$

where we used the assumption that $\mathbf{X}$ is conditionally independent of $\mathbf{\Pi}$ given $\mathbf{Z}$ and $\mathbf{Z}$ is independent of $\mathbf{U}$. Here, $\mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_I]$ denotes the parameters in I Dirichlet mixture components. The conditional PDF of $\mathbf{X}$ given $\{\mathbf{Z},\mathbf{U}\}$ can be written as

$$f(\mathbf{X}|\mathbf{Z},\mathbf{U}) = \prod_{n=1}^{N}\prod_{i=1}^{I}[\mathrm{Dir}(\mathbf{x}_n|\mathbf{u}_i)]^{z_{ni}}. \tag{6}$$

Assuming that the indication vectors $\mathbf{Z}$ are random vector variables, the conditional PDF of $\mathbf{Z}$ given $\mathbf{\Pi}$ is categorial distributed as

$$f(\mathbf{Z}|\mathbf{\Pi}) = \prod_{n=1}^{N}\prod_{i=1}^{I}\pi_i^{z_{ni}}. \tag{7}$$

Also, the prior distribution of $\mathbf{\Pi}$ is assumed to be Dirichlet distributed as

$$f(\mathbf{\Pi}) = \mathrm{Dir}(\mathbf{\Pi};\mathbf{c_0}), \tag{8}$$

where $\mathbf{c}_0=[c_{1_0},\ldots,c_{I_0}]^{\mathrm{T}}$ is the hyperparameter vector in the prior distribution of $\mathbf{\Pi}$. If we take the prior distribution of $\mathbf{u}_i$ as the conjugate prior in (3), an analytically tractable solution cannot be obtained. With the principles of the VI framework [18,3], we assume that the $K+1$ variables in $\mathbf{u}_i$ are mutually independent, factorize them into disjoint groups (one group has only one variable), and assign a gamma distribution to each variable as the prior distribution. Then the prior joint distribution of $\mathbf{U}$ is

$$f(\mathbf{U}) = \prod_{i=1}^{I}\prod_{k=1}^{K+1}\mathrm{Gam}(u_{ki};\mu_{ki_0},\alpha_{ki_0}), \tag{9}$$

where μ_{ki_0} and α_{ki_0} are the hyperparameters in the prior distribution of u_{ki} and the PDF of the gamma distribution is defined as

$$\mathrm{Gam}(u;\mu,\alpha) = \frac{\alpha^{\mu}}{\Gamma(\mu)}u^{\mu-1}e^{-\alpha u}. \tag{10}$$

Fig. 1 illustrates the variables' relations by a graphic model. With Bayes' theorem and combining (5)–(9) together, we can represent the joint distribution of the observation $\mathbf{X}$ and all the latent variables $\mathcal{Z}=\{\mathbf{U},\mathbf{\Pi},\mathbf{Z}\}$ by

$$\begin{aligned} f(\mathbf{X},\mathcal{Z}) &= f(\mathbf{X},\mathbf{U},\mathbf{\Pi},\mathbf{Z}) \\ &= f(\mathbf{X}|\mathbf{Z},\mathbf{U})f(\mathbf{Z}|\mathbf{\Pi})f(\mathbf{\Pi})f(\mathbf{U}) \\ &= \prod_{n=1}^{N}\prod_{i=1}^{I}\left[\pi_i\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})}\prod_{k=1}^{K+1}x_{kn}^{u_{ki}-1}\right]^{z_{ni}} \times \frac{\Gamma(\sum_{i=1}^{I}c_{i_0})}{\prod_{i=1}^{I}\Gamma(c_{i_0})}\prod_{i=1}^{I}\pi_i^{c_{i_0}-1} \\ &\quad \times \prod_{i=1}^{I}\prod_{k=1}^{K+1}\frac{\alpha_{ki_0}^{\mu_{ki_0}}}{\Gamma(\mu_{ki_0})}u_{ki}^{\mu_{ki_0}-1}e^{-\alpha_{ki_0}u_{ki}}. \end{aligned} \tag{11}$$

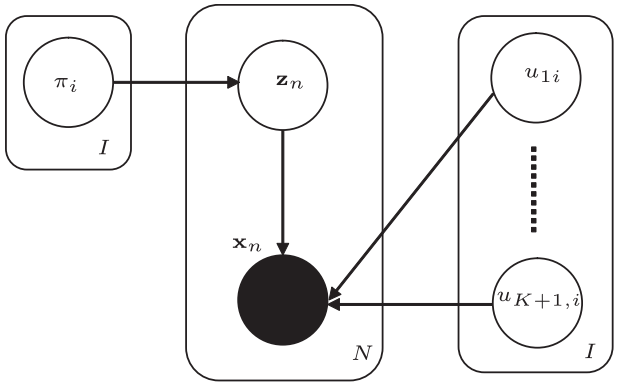

Fig. 1. Graphical representation of the variables' relationship in the Bayesian inference of a DMM, where the parameters in a Dirichlet distribution are assumed to be mutually independent. All the circles in the graphical figure represent variables. Arrows show the relationship between variables.

In the VI framework [3,18–20], the variational objective function (variational lower-bound) that we would like to maximize is

$$\mathcal{L} = \mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X},\mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})]. \tag{12}$$

This is actually the Kullback–Leibler (KL) divergence from the joint distribution $f(\mathbf{X},\mathcal{Z})$ to the approximating posterior distribution $f(\mathcal{Z})$. According to the factorized approximation (FA) method [3, Chapter 10], if we consider that all the variables in $\mathcal{Z}$ are mutually independent and update only one variable by fixing the others' distributions for a moment, the optimal solution to the posterior distribution of u_{ki} is

$$\begin{aligned} &\ln f^*(u_{ki};\mu_{ki},\alpha_{ki}) \\ &\quad = \mathbb{E}_{\mathcal{Z}\setminus u_{ki}}[\ln f(\mathbf{X},\mathcal{Z})] \\ &\quad = \mathbb{E}_{\mathcal{Z}\setminus u_{ki}}\left\{\sum_{n=1}^{N}\sum_{i=1}^{I}z_{ni}\left[\ln\pi_i + \ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})} + \sum_{k=1}^{K+1}(u_{ki}-1)\ln x_{kn}\right]\right. \\ &\qquad + \ln\frac{\Gamma(\sum_{i=1}^{I}c_{i_0})}{\prod_{i=1}^{I}\Gamma(c_{i_0})} + \sum_{i=1}^{I}(c_{i_0}-1)\ln\pi_i \\ &\qquad \left. + \sum_{i=1}^{I}\sum_{k=1}^{K+1}\left[\ln\frac{\alpha_{ki_0}^{\mu_{ki_0}}}{\Gamma(\mu_{ki_0})} + (\mu_{ki_0}-1)\ln u_{ki} - \alpha_{ki_0}u_{ki}\right]\right\}, \end{aligned} \tag{13}$$

where $\mathbb{E}_{\mathcal{Z}\setminus u_{ki}}$ denotes expectation with respect to the variational distribution of all the variables except for u_{ki}. When the expectations with respect to the other variables are absorbed in the constant term, the optimal solution is

$$\begin{aligned} &\ln f^*(u_{ki};\mu_{ki},\alpha_{ki}) \\ &\quad = \sum_{n=1}^{N}\mathbb{E}[z_{ni}] \times \mathbb{E}_{\mathcal{Z}\setminus u_{ki}}\left[\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})}\right] \\ &\qquad + u_{ki}\sum_{n=1}^{N}\mathbb{E}[z_{ni}]\ln x_{kn} + (\mu_{ki_0}-1)\ln u_{ki} - \alpha_{ki_0}u_{ki} + \mathrm{const.} \end{aligned} \tag{14}$$

When π_i is the random variable to be updated, the optimal solution to the posterior distribution of π_i is

$$\ln f^*(\pi_i;c_i) = \mathbb{E}_{\mathcal{Z}\setminus\pi_i}[\ln f(\mathbf{X},\mathcal{Z})] = \ln\pi_i \times \sum_{n=1}^{N}\mathbb{E}[z_{ni}] + \ln\pi_i(c_{i_0}-1) + \mathrm{const.} \tag{15}$$

Similarly, for the variable z_{ni}, the optimal approximation to the posterior distribution is

$$\begin{aligned} \ln f^*(z_{ni}) &= \mathbb{E}_{\mathcal{Z}\setminus z_{ni}}[\ln f(\mathbf{X},\mathcal{Z})] \\ &= z_{ni} \times \left[\mathbb{E}[\ln\pi_i] + \sum_{k=1}^{K+1}(u_{ki}-1)\ln x_{kn}\right] \\ &\quad + z_{ni} \times \mathbb{E}\left[\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})}\right] + \mathrm{const.} \end{aligned} \tag{16}$$

If the optimal solutions in (14)–(16) have the same form as the logarithm of their prior distributions, we can obtain an analytically tractable solution for the Bayesian analysis of DMM. This is one advantage of the FA method. It is also the main reason that we factorized the jointly distributed parameters into disjoint groups. However, in the above three optimal solutions, only (15) has the same form as the logarithm of the Dirichlet prior distribution. The remaining two optimal posterior distributions do not have the same logarithm as their prior distributions. Thus, we cannot obtain an analytically tractable solution. In the next sections, we use the extended factorized approximation (EFA) method (see *e.g.*, [27,23,24,8]) to derive an analytically tractable solution.

3.3. Extended factorized approximation with single lower-bound approximation

By the factorized approximation (FA) method in the VI framework [3,18–20], the variational objective function that we would

like to maximize is

$$\mathcal{L} = \mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})]. \tag{17}$$

As mentioned above, we cannot obtain an analytically tractable solution directly with the FA method. This is because the optimal solution depends on the expectation computed with respect to the factor distribution [3]. Therefore, some approximations can be applied to get a nearly optimal analytically tractable solution. This theorem is the so-called extended factorized approximation (EFA) method [18,8]. Braun et al. [23] considered the zeroth-order and the first-order delta method for moments [28] to derive an alternative for the objective function to simplify the calculation. Blei and Lafferty [21,22] proposed a correlated topic model (CTM) and used a first-order Taylor expansion to preserve a bound such that an intractable expectation was avoided. A similar idea was also applied in [8] for approximating the posterior distributions in BMM. Using Jensen's inequality has become commonplace in variational inference. With Jensen's inequality, the posterior distribution can be approximated by a tractable Gaussian [27]. In [24], the concavity of the function $-x^{-1}$ and the convexity of $-\log x$ were studied and Jensen's inequality and the first-order Taylor expansion were applied to approximately calculate the posterior distribution. The above-mentioned methods were actually using the following property:

If we could find an (unnormalized) likelihood function $\widetilde{f}(\mathbf{X}, \mathcal{Z})$ which satisfies

$$\mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})] \geq \mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})], \tag{18}$$

then the variational objective function (see [3, p. 465]) can be lower-bounded as

$$\mathcal{L} = \mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})] \geq \mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})]. \tag{19}$$

Even though we cannot maximize $\mathcal{L}$ directly, by maximizing the lower-bound of $\mathcal{L}$, we can still reach the maximum value of $\mathcal{L}$ asymptotically [18,27,24,8]. The new bound is tight at one point from the parameter distribution [8].

This property is the so-called extended factorized approximation (EFA) method [18,21,22,8]. By the EFA method and for the purpose of conjugate match, if we could find a auxiliary function $\widetilde{f}(\mathbf{X}, \mathcal{Z})$ which satisfies (18), then we can maximize

$$\widetilde{\mathcal{L}} = \mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})], \tag{20}$$

which is a lower-bound to the variational objective function in (17), to get an analytically tractable solution. The approximated optimal solution in this case is

$$\ln \widetilde{f}_i^*(\mathcal{Z}_i) = \mathbb{E}_{\mathcal{Z}\setminus\mathcal{Z}_i}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})] + \text{const}. \tag{21}$$

Instead of maximizing the original object function directly, maximizing a lower-bound is equivalent to maximizing the original variational objective function asymptotically. Although introducing a systematic gap when involving the lower-bound approximation, the EFA allows flexibility when dealing with intractable integrations and provides a convenient way to obtain an analytically tractable solution. Similar ideas were also applied in, *e.g.*, [18,21–24].

For a DMM, the difficulty in calculating the expectation of the first term comes from the multivariate log-inverse-beta (MLIB) function $\ln \Gamma(\sum_{k=1}^{K+1} u_{ki}) / \prod_{k=1}^{K+1} \Gamma(u_{ki})$ in $\ln f(\mathbf{X}, \mathcal{Z})$ (see (11)). Now we study the relative convexity of this MLIB function to find a lower-bound approximation to it.

Theorem 1. *The MLIB function* $\ln \Gamma(\sum_{k=1}^{K+1} u_{ki}) / \prod_{k=1}^{K+1} \Gamma(u_{ki})$ *is convex relative to* $\ln u_{ki}$ *[29] when* $\sum_{m \neq k, m=1}^{K+1} u_{mi} > 1$.[2] *Thus, this MLIB function can be lower-bounded by its first-order Taylor expansion for* $\ln u_{ki}$.

Proof of Theorem 1 can be found in Appendix A. With this theorem, a lower-bound approximation to the MLIB function can be obtained as

$$\ln \frac{\Gamma(\sum_{k=1}^{K+1} u_{ki})}{\prod_{k=1}^{K+1} \Gamma(u_{ki})} \geq \ln \frac{\Gamma(\sum_{k=1}^{K+1} \overline{u}_{ki})}{\prod_{k=1}^{K+1} \Gamma(\overline{u}_{ki})} + \sum_{k=1}^{K+1} \left[\psi\left(\sum_{m=1}^{k} \overline{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) - \psi(\overline{u}_{ki}) \right] \overline{u}_{ki} (\ln u_{ki} - \ln \overline{u}_{ki}), \tag{22}$$

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

Substituting (22) into (17), the first term of the variational objective function, *i.e.*, $\mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})]$, is lower-bounded as

$$\begin{aligned} \mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})] \geq & \sum_{n=1}^{N} \sum_{i=1}^{I} \mathbb{E}[z_{ni}] \left(\ln \frac{\Gamma(\sum_{k=1}^{K+1} \overline{u}_{ki})}{\prod_{k=1}^{K+1} \Gamma(\overline{u}_{ki})} \right. \\ & + \sum_{k=1}^{K+1} \mathbb{E}_{u_{p \geq k,i}} \left\{ \left[\psi\left(\sum_{m=1}^{k} \overline{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) - \psi(\overline{u}_{ki}) \right] \right. \\ & \left. \left. \times \overline{u}_{ki} (\ln u_{ki} - \ln \overline{u}_{ki}) \right\} \right) + \mathcal{R}^{\dagger} \end{aligned}$$

$^{\dagger}\mathcal{R}$ denotes the remaining unchanged parts in $\mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X}, \mathcal{Z})]$. (23)

The second expectation with respect to $u_{p \geq j,i}$ is not directly feasible. With Jensen's inequality, we have the following two inequalities:

$$\mathbb{E}_x[\psi(x+y)] \leq [\psi(\overline{x}+y)], \quad \mathbb{E}_x[\ln x] \leq \ln \overline{x}, \tag{24}$$

as both $\psi(x+y)$ and $\ln x$ are concave functions in x. Then the second expectation can be lower-bounded as

$$\begin{aligned} & \mathbb{E}_{u_{p \geq k,i}} \left\{ \left[\psi\left(\sum_{m=1}^{k} \overline{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) - \psi(\overline{u}_{ki}) \right] \overline{u}_{ki} (\ln u_{ki} - \ln \overline{u}_{ki}) \right\} \\ & = \left\{ \mathbb{E}_{u_{p>k,i}} \left[\psi\left(\sum_{m=1}^{k} \overline{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) \right] - \psi(\overline{u}_{ki}) \right\} \overline{u}_{ki} \left(\underbrace{\mathbb{E}_{u_{ki}}[\ln u_{ki}] - \ln \overline{u}_{ki}}_{\leq 0} \right) \\ & \geq \left[\psi\left(\sum_{m=1}^{k} \overline{u}_{mi} + \mathbb{E}_{u_{p>k,i}} \left[\sum_{l=k+1}^{K+1} u_{li} \right] \right) - \psi(\overline{u}_{ki}) \right] \overline{u}_{ki} (\mathbb{E}_{u_{ki}}[\ln u_{ki}] - \ln \overline{u}_{ki}) \\ & = \left[\psi\left(\sum_{k=1}^{K+1} \overline{u}_{ki} \right) - \psi(\overline{u}_{ki}) \right] \overline{u}_{ki} (\mathbb{E}_{u_{ki}}[\ln u_{ki}] - \ln \overline{u}_{ki}) \end{aligned} \tag{25}$$

Finally, the variational objective function $\mathcal{L}$ in (17) is approximated by a single lower-bound (SLB) as

$$\mathcal{L} \geq \widetilde{\mathcal{L}} = \mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})] - \mathbb{E}_{\mathcal{Z}}[\ln f(\mathcal{Z})], \tag{26}$$

where $\mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})]$ is represented as

$$\begin{aligned} & \mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})] \\ & = \sum_{n=1}^{N} \sum_{i=1}^{I} \mathbb{E}[z_{ni}] \left\{ \ln \frac{\Gamma(\sum_{k=1}^{K+1} \overline{u}_{ki})}{\prod_{k=1}^{K+1} \Gamma(\overline{u}_{ki})} \right. \\ & \left. + \sum_{k=1}^{K+1} \left[\psi\left(\sum_{k=1}^{K+1} \overline{u}_{ki} \right) - \psi(\overline{u}_{ki}) \right] \overline{u}_{ki} (\mathbb{E}_{u_{ki}}[\ln u_{ki}] - \ln \overline{u}_{ki}) \right\} + \mathcal{R}. \end{aligned} \tag{27}$$

Note that $\mathbb{E}_{\mathcal{Z}}[\ln \widetilde{f}(\mathbf{X}, \mathcal{Z})]$ satisfies (18). Thus maximizing $\widetilde{\mathcal{L}}$ is asymptotically equivalent to maximizing $\mathcal{L}$.

3.4. Tightness of the lower-bound

In the above derivations, we used two approximations, namely the first-order Taylor expansion and Jensen's inequality. For the first-order Taylor expansion in (22), we expanded the MLIB function around the logarithmic mean values of the arguments.

[2] As mentioned in Section 2, we only consider the case where each of the parameters in a Dirichlet distribution is greater than one, which means that *any* parameter has almost zero probability of taking the values smaller than one.

In (24), Jensen's inequality was applied. By these two approximations, the expectation of the MLIB function is lower-bounded. In practice, choosing the logarithmic mean value as the expansion point and applying Jensen's inequality can actually tighten the lower-bound. The tightness of these approximations is discussed in Appendix B.

3.5. An analytically tractable solution

By the relative convexity, we derived SLB approximations to the variational objective function $\mathcal{L}$ in the previous section. With these lower-bound approximations, the variational objective function $\mathcal{L}$ is lower-bounded by $\widetilde{\mathcal{L}}$. Instead of maximizing the variational objective function directly, we can maximize $\widetilde{\mathcal{L}}$ to get an approximation to the optimal solution [24,8]. With the principle of the VI framework, the optimal solution to the posterior distribution of u_{ki} is[3]

$$\begin{aligned}\ln f^*(u_{ki};\mu_{ki},\alpha_{ki}) &\approx \mathbb{E}_{\mathcal{Z}\setminus u_{ki}}[\ln \widetilde{f}(\mathbf{X},\mathcal{Z})]\\ &= \sum_{n=1}^{N}\mathbb{E}[z_{ni}]\times\left[\psi\left(\sum_{k=1}^{K+1}\overline{u}_{ki}\right)-\psi(\overline{u}_{ki})\right]\overline{u}_{ki}(\ln u_{ki}-\ln\overline{u}_{ki})\\ &\quad+u_{ki}\sum_{n=1}^{N}\mathbb{E}[z_{ni}]\ln x_{kn}+(\mu_{ki_0}-1)\ln u_{ki}-\alpha_{ki_0}u_{ki}+\text{const.},\end{aligned}\tag{28}$$

which has the logarithmic form of the gamma distribution. For the same reason, the optimal solution to the posterior distribution of z_{ni} is

$$\begin{aligned}\ln f^*(z_{ni}) &\approx \mathbb{E}_{\mathcal{Z}\setminus z_{ni}}[\ln f(\mathbf{X},\mathcal{Z})]\\ &= z_{ni}\times\left[\mathbb{E}[\ln\pi_i]+\sum_{k=1}^{K+1}(u_{ki}-1)\ln x_{kn}\right]\\ &\quad+z_{ni}\times\underbrace{\sum_{k=1}^{K+1}\left[\psi\left(\sum_{k=1}^{K+1}\overline{u}_{ki}\right)-\psi(\overline{u}_{ki})\right]\overline{u}_{ki}(\mathbb{E}_{u_{ki}}[\ln u_{ki}]-\ln\overline{u}_{ki})}_{P_i}+\text{const.}\end{aligned}\tag{29}$$

Similarly, it is the logarithmic form of the categorial distribution. The analytically tractable solution for Bayesian estimation of a DMM is summarized in Algorithm 1, where the required quantities are

$$\overline{u}_{ki}=\frac{\mu_{ki}}{\alpha_{ki}},\quad \mathbb{E}[z_{ni}]=\frac{\rho_{ni}}{\sum_{i=1}^{I}\rho_{ni}},$$
$$\ln\rho_{ni}=\psi(c_i)-\psi(\mathbf{c}^{\mathrm{T}}\mathbf{1}_I)+P_i+(\mathbf{u}_i-1)^{\mathrm{T}}\ln\mathbf{x}_n.\tag{30}$$

This analytically tractable solution avoids the numerical calculation used in the EM algorithm [30], hence the calculation is facilitated. Fig. 2 shows the convergence of the proposed algorithm. The true variational objective function was calculated numerically by the importance sampling method during each iteration. It can be observed that the variational objective function was always increasing during iterations, hence the convergence is demonstrated.

Algorithm 1. Variational DMM.

Input: observation $\mathbf{X}$, number of mixture components I
Initialize α_{ki_0}, μ_{ki_0}, c_{i_0}, for $i=1,\ldots,I$, $k=1,\ldots,K+1$[3];
repeat
for each k, i
$\alpha^*_{ki}=\alpha_{ki_0}-\sum_{n=1}^{N}\mathbb{E}[z_{ni}]\ln x_{kn}$
$\mu^*_{ki}=\mu_{ki_0}+\sum_{n=1}^{N}\mathbb{E}[z_{ni}]\overline{u}_{ki}[\psi(\sum_{k=1}^{K+1}\overline{u}_{ki})-\psi(\overline{u}_{ki})]$
$c^*_i=c_{i_0}+\sum_{n=1}^{N}\mathbb{E}[z_{ni}]$

[3] In practice, from a Bayesian perspective, we can set all the gamma priors to have the same prior hyperparameters.

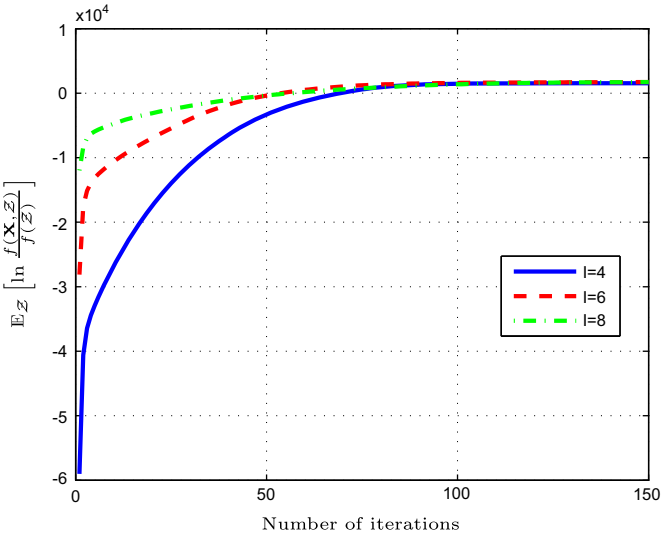

Fig. 2. Convergence of the proposed Bayesian estimation method. The observation $\mathbf{X}$ with 1000 vectors was generated from a DMM with $\pi_1=0.35$, $\mathbf{u}_1=[4\ 12\ 3]^{\mathrm{T}}$ and $\pi_2=0.65$, $\mathbf{u}_2=[10\ 6\ 2]^{\mathrm{T}}$. We ran the proposed algorithm with different initial mixture components, *i.e.*, $I=4, 6, 8$, and they all kept two mixture components after convergence. At different initial settings, the algorithm converges after 70–120 iterations. For some other parameter settings, similar performances can also be obtained. We show only one example here. The variational objective function was calculated numerically.

until stop criteria are reached.
Output: the optimal hyperparameters α^*_{ki}, μ^*_{ki}, c^*_i.
(The quantities $\overline{u}_{ki}$ and $\mathbb{E}[z_{ni}]$ are calculated according to (30).)

When we get the posterior distribution from Algorithm 1, the point estimates to the DMM parameters can be obtained by taking the posterior means of the gamma distributions, which are

$$\hat{u}_{ki}=\frac{\mu^*_{ki}}{\alpha^*_{ki}},\quad k=1,\ldots,K,\ \ i=1,\ldots,I.\tag{31}$$

Meanwhile, the point estimates to the mixture weights are

$$\hat{\pi}_i=\frac{c^*_i}{\sum_{i=1}^{I}c^*_i},\quad i=1,\ldots,I.\tag{32}$$

3.6. Discussion

With the above-derived algorithm, the posterior distributions of the parameters in a DMM can be calculated approximately. Also, we can get the point estimates to the parameters in the DMM by (31) and (32). In this section, we discuss two important issues in the proposed VI-based method: (1) the systematic bias and (2) the convergency.

3.6.1. Systematic bias

The VI framework could introduce some systematic bias by factorizing the parameters into disjoint groups. However, this effect is negligible, compared to the advantage of the analytically tractable solution. The proposed algorithm is guaranteed to converge, since the variational objective function $\mathcal{L}$ is lower-bounded by $\widetilde{\mathcal{L}}$ and maximized asymptotically. But it may converge to a local maximum because the variational objective function is convex in each parameter [3, Chapter 10] but not jointly convex for all the parameters. Compared to some numerical sampling method (*e.g.*, Markov chain Monte Carlo) based approach, the proposed Bayesian approach incurs unknown bias [4] because of the lower-bound approximation. But the Bayesian approach is deterministic

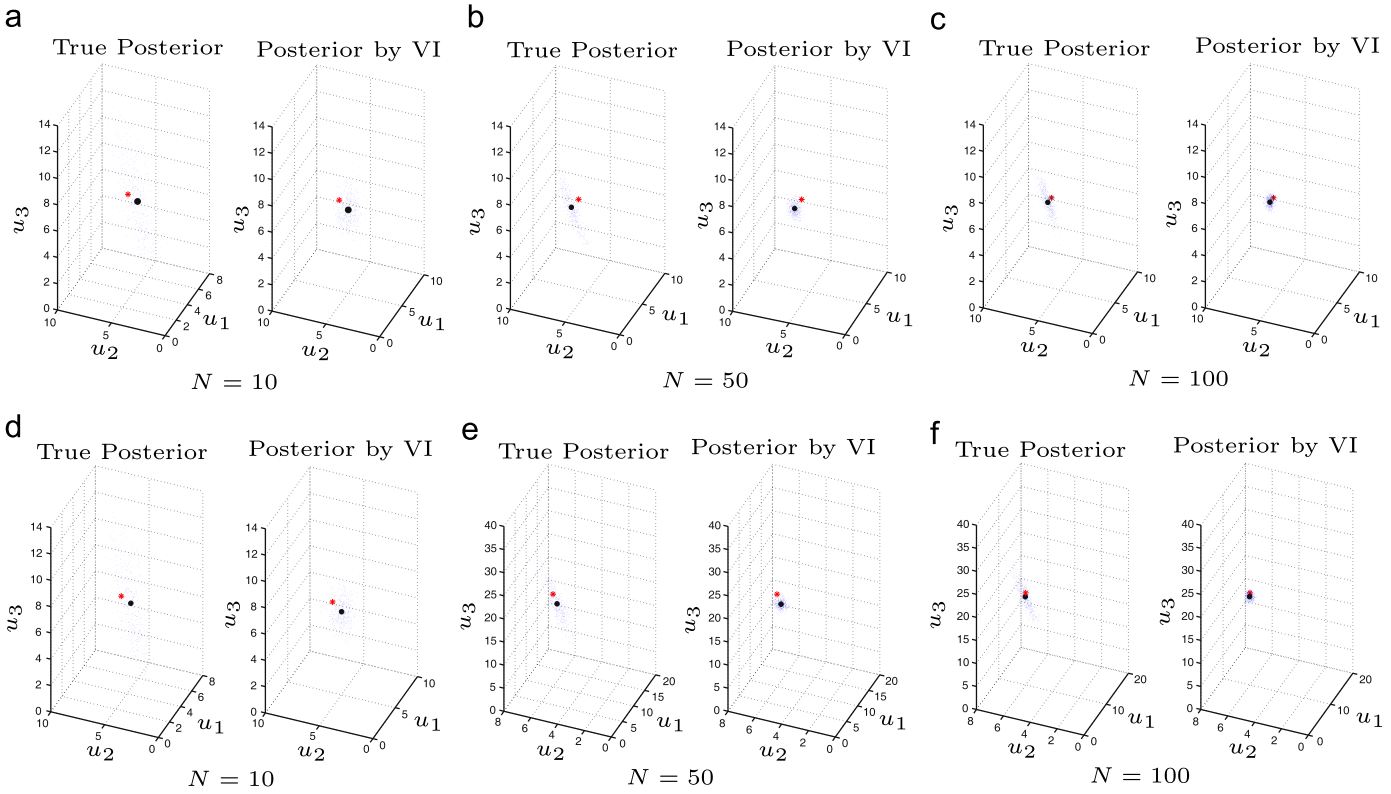

Fig. 3. Comparisons of the true posterior distribution and the approximating one obtained by VI. The true posterior distribution was obtained by the rejection sampling method [3], where the referred distribution is the Laplace approximation to the true one. Different amounts of training data were generated from the Dirichlet distribution with known $\mathbf{u}$. The red star shows the true parameter. The black dot is the posterior mean in either the true posterior distribution or the posterior distribution obtained by VI. (a)–(c) show the comparison with $\mathbf{u}=[3\ 5\ 8]^{\mathrm{T}}$. (d)–(f) show the comparison with $\mathbf{u}=[10\ 6\ 20]^{\mathrm{T}}$. It can be observed that both the true and the approximating posterior distribution can estimate the parameters properly. Meanwhile, the mismatch between the true and the approximating posterior distribution is also illustrated, which is due to the assumption of mutual independence of the parameters in the Dirichlet distribution. However, the difference becomes smaller as the amount of data increases. As expected [3], the VI-based method yields a distribution in a more compact parameter space than the true one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

and is guaranteed to converge. Since the sampling can be a slow process for convergence, we prefer the Bayesian approach to the sampling methods, even though some bias may be incurred. When we have a large amount of observed data, the posterior distributions, which are unimodally distributed, will become narrow and concentrate around the mean value. The accuracy of the proposed method will be increased (see Fig. 3).

3.6.2. Convergence of the EFA method

In this paper, we used *single* lower-bound (SLB) approximation (see (26)) to derive an analytically tractable solution. The variational objective function is lower-bounded by a SLB $\widetilde{\mathcal{L}}$ and this lower-bound is maximized during each iteration. As a single function is maximized iteratively, the convergence is guaranteed.

A two-dimensional Dirichlet distribution (which has the degrees of freedom equal to $K=1$) is actually a beta distribution. If we set $K=1$, the proposed Bayesian estimation method should be identical to our Bayesian estimation method of the BMM proposed in [8]. However, it is *not*. This is because the proposed method in [8] utilized *multiple* lower-bounds (MLB) approximation to the variational objective function.

In [8], we approximated the log-inverse-beta (LIB) function $\ln\Gamma(u+v)/\Gamma(u)\Gamma(v)$ by *three* different lower-bounds, one with respect to u [8, Eq. (26)], the other one with respect to v [8, Eq. (27)], and the third one with respect to $[u,v]^T$ [8, Eq. (28)]. With these approximations, we maximized each lower-bound separately and obtained an analytically tractable solution. However, *this is an intuitively correct but actually not accurate strategy*.

Let us first take a simple case with two different lower-bounds. Assuming that the random variable set $\mathbf{Z}$ is divided into two disjoint variable groups $\mathbf{Z}=\{\mathbf{Z}_1,\mathbf{Z}_2\}$ and for each random variable group, we have two different auxiliary functions, *e.g.*, obtained by the Taylor expansion, $\widetilde{f}_1(\mathbf{X},\mathbf{Z})$[4] with respect to $\mathbf{Z}_1$ and $\widetilde{f}_2(\mathbf{X},\mathbf{Z})$ with respect to $\mathbf{Z}_2$, which satisfy

$$\mathbb{E}_{\mathbf{Z}}[\ln f(\mathbf{X},\mathbf{Z})]\geq\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_1(\mathbf{X},\mathbf{Z})],\quad \mathbb{E}_{\mathbf{Z}}[\ln f(\mathbf{X},\mathbf{Z})]\geq\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_2(\mathbf{X},\mathbf{Z})]. \tag{33}$$

Then with EFA, we have two lower-bounds as

$$\widetilde{\mathcal{L}}_1=\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_1(\mathbf{X},\mathbf{Z})-\ln f(\mathbf{Z})],\quad \widetilde{\mathcal{L}}_2=\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_2(\mathbf{X},\mathbf{Z})-\ln f(\mathbf{Z})]. \tag{34}$$

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

$$\ln\widetilde{f}_1^*(\mathbf{Z}_1)=\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_1}[\ln\widetilde{f}_1(\mathbf{X},\mathbf{Z})]+\text{const.} \tag{35a}$$

$$\ln\widetilde{f}_2^*(\mathbf{Z}_2)=\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_2}[\ln\widetilde{f}_2(\mathbf{X},\mathbf{Z})]+\text{const.} \tag{35b}$$

With the above strategy, it looks like what we are maximizing is just two times the original lower-bound as

$$2\times\mathcal{L}\geq\widetilde{\mathcal{L}}_1+\widetilde{\mathcal{L}}_2=\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_1(\mathbf{X},\mathbf{Z})]-\mathbb{E}_{\mathbf{Z}}[\ln f(\mathbf{Z})] \tag{36a}$$

$$2\times\mathcal{L}\geq\widetilde{\mathcal{L}}_1+\widetilde{\mathcal{L}}_2=+\mathbb{E}_{\mathbf{Z}}[\ln\widetilde{f}_2(\mathbf{X},\mathbf{Z})]-\mathbb{E}_{\mathbf{Z}}[\ln f(\mathbf{Z})]. \tag{36b}$$

[4] The auxiliary function in terms of $\mathbf{Z}_1$ is also a function of the expansion point $\widetilde{\mathbf{Z}}_1$. We omit it in the expression and keep only the random variables $\mathbf{Z}=\{\mathbf{Z}_1,\mathbf{Z}_2\}$ there.

When performing the update strategy (35a), we get (36a) to be maximized. This maximization makes the distribution of $\mathbf{Z}_1$ to be less uncertain. As $-\mathbb{E}_{\mathbf{Z}}[\ln f(\mathbf{Z})]$ in (36b) is the differential entropy of $\mathbf{Z}$, (36b) is decreasing while (36a) is maximizing. It is hard to evaluate if (36a) changes more than (36b) or not. Thus, the overall lower-bound, *i.e.*, $\widehat{\mathcal{L}}_1+\widehat{\mathcal{L}}_2$, might decrease during some iterations. On the one hand, as the lower-bound to the variational objective function cannot be guaranteed to be maximized all the time, this strategy may not promise convergence. On the other hand, if the change to (36a) is larger than (36b), the convergence is still guaranteed. The same argument applies to the EFA with more than two lower-bounds. Thus, the convergence of the algorithm proposed in [8] is underdetermined, although we can show the convergence empirically with simulations.

As the Dirichlet distribution is the extended version of the beta distribution, similar arguments are also valid for the DMM case. In [12], the authors derived different lower-bounds for different variables when deriving the VI based solution for the DMM. Similar to the variational estimation of the BMM in [8], they also used the MLB approximation strategy ($K+2$ lower-bounds for a Dirichlet distribution with $K+1$ dimensions) strategy and maximized each lower-bound separately. As argued above, since there is no single objective function to be maximized during the iterations, the convergence of the proposed method in [12] is not guaranteed either.

4. Experimental results and discussion

We evaluate the proposed method with both synthesized data and real data. With the synthesized data, the efficiency and accuracy of the proposed Bayesian estimation method are demonstrated. In the real data evaluation, the proposed Bayesian DMM method is applied for the purpose of LSF parameters quantization in speech transmission and compared to the recently proposed statistical model based methods [13]. Meanwhile, we also apply the proposed Bayesian DMM to enhance the multiview depth image, which plays a pivotal role in the emerging free-viewpoint television.

In principle, the Dirichelt distribution is devoted to negatively correlated data [31]. Hence, before applying the Dirichlet distribution, the data correlation should be checked. In our experiments, the synthesized data are generated from the Dirichlet distribution/DMM. Thus, they are negatively correlated. For the real data evaluation, the ΔLSF parameters and the three-dimensional chromaticity value in the XYZ space were shown to be negatively correlated in [9] and [32], respectively.

In the initialization phase of all the experiments, we set the hyperparameters of the prior distribution as $\alpha_{ki_0}=0.001$, $\mu_{ki_0}=1$, and $c_{i_0}=0.0001$ for $k=1,\ldots,K+1$ and $i=1,\ldots,I$, which mean that all the gamma priors have the same hyperparameters and all the mixture components have the same initial weights. The prior distributions are therefore non-informative. Moreover, this setting is according to our experimental experiences. It is worthy to note that, in such a case, the prior means of the Dirichlet parameters are far greater than 1 ($\overline{u}_{ki_0}=\mu_{ki_0}/\alpha_{ki_0}$). Thus, the constraint mentioned in Section 2 is satisfied. After the algorithm converges, we take the posterior means as the point estimates to the parameters in a DMM.

4.1. Synthesized data evaluation

In this section, we evaluate the proposed Bayesian estimation method with several synthesized data sets. The bias and the accuracy of the proposed method are demonstrated by the synthesized data evaluation.

4.1.1. Bias of the proposed method

As mentioned in Section 3.6, the assumption of the mutual independence among the parameters in the Dirichlet distribution violates the actual correlation. In Fig. 3, we compare the true posterior distribution with the approximating one obtained by the proposed VI-based method. It can be observed that there is a mismatch between the true and the approximation. However, as the amount of observation increases, both the true and the approximating posterior distribution tend to concentrate in a narrow area, thus the difference between these two can be neglected. Furthermore, it is reasonable to take the posterior mean as the point estimate when the amount of observation is sufficiently large. Here, we use a single three-dimensional Dirichlet distribution purely for the purpose of a convenient visualization. More examples with more mixture components and higher dimensional data are shown in Section 4.1.2.

4.1.2. Efficiency of mixture modeling

Another advantage of the Bayesian estimation over the ML estimation is that it could estimate the model complexity properly based on the observed data. This task is challenging in model estimation and the ML estimation itself cannot handle this issue. In the Bayesian estimation, a mixture component which contains extremely a small amount of data will be discarded from the whole model by assigning zero (or a small number that is undistinguished from zero) probability to it. Thus the overfitting problem which is a drawback of the ML estimation can be avoided. Table 1 shows examples of estimating the model complexity with the proposed Bayesian method. For each data set generated from a DMM, the Bayesian method was initialized with double amounts of the true number of mixture components. It can be observed that the Bayesian method can estimate the model complexity properly. When the number of observations increases, the point estimates obtained from the Bayesian method become more accurate. The posterior means could be used as reasonable and reliable point estimates to the true parameters.

In the case of Dirichlet distribution, the data with modes close to 0 or 1 are generally difficult to estimate. To check the robustness of the proposed VI-based method, we applied it to estimate the underlying distribution of such skewed distributed data and study the efficiency in this case. Table 2 shows an example. The distribution with this parameter setting has two modes which are all close to 0 or 1. The proposed VI-based method can estimate the distribution properly.

4.1.3. Comparisons with the ML estimation

With the observed data, the parameters in the model can be estimated. In addition to the estimated parameters, the ultimate goal in statistical modeling is to get a predictive distribution, with which the likelihood of a new data $\mathbf{x}$ given the observed ones $\mathbf{X}$ can be predictively calculated. The predictive distribution formally writes

$$f(\mathbf{x}|\mathbf{X})=\int \mathrm{Dir}(\mathbf{x}|\mathbf{u})f(\mathbf{u}|\mathbf{X})\,d\mathbf{u}. \tag{37}$$

However, in most cases, the predictive distribution cannot be calculated by an analytically tractable solution. With the obtained point estimates, we can approximate the predictive distribution as

$$f(\mathbf{x}|\mathbf{X})\approx f(x;\widehat{\mathbf{u}})=\mathrm{Dir}(\mathbf{x};\widehat{\mathbf{u}}), \tag{38}$$

where $\widehat{\mathbf{u}}$ denoted the estimated parameters. One advantage of the Bayesian estimation over the conventional ML estimation is

Table 1
Efficiency of the proposed Bayesian estimation method with synthesized data evaluations. At initialization, the number of mixture components was set to be two times of the true one. After convergence, the correct number of mixture components was estimated. The components with extremely small weights were discarded and not listed in the table.

	Estimated parameters	Estimated parameters
	Model A: $\pi_1=0.65,\ \mathbf{u}_1=[4\ 12\ 3]^T$ $\pi_2=0.35,\ \mathbf{u}_2=[10\ 6\ 2]^T$	**Model B**: $\pi_1=0.2,\ \mathbf{u}_1=[3\ 5\ 12\ 6]^T$ $\pi_2=0.5,\ \mathbf{u}_2=[4\ 12\ 3\ 9]^T$ $\pi_3=0.3,\ \mathbf{u}_3=[10\ 6\ 2\ 5]^T$
$N=100$	$\hat{\pi}_1=0.347,\ \hat{\mathbf{u}}_1=[13.9\ 8.27\ 2.55]^T$ $\hat{\pi}_2=0.653,\ \hat{\mathbf{u}}_2=[4.90\ 13.00\ 3.70]^T$	$\hat{\pi}_1=0.210,\ \hat{\mathbf{u}}_1=[3.45\ 5.36\ 15.15\ 7.01]^T$ $\hat{\pi}_4=0.282,\ \hat{\mathbf{u}}_4=[12.24\ 6.93\ 2.08\ 6.71]^T$ $\hat{\pi}_5=0.508,\ \hat{\mathbf{u}}_5=[4.08\ 12.52\ 3.29\ 9.36]^T$
$N=200$	$\hat{\pi}_1=0.655,\ \hat{\mathbf{u}}_1=[4.06\ 13.11\ 2.92]^T$ $\hat{\pi}_3=0.345,\ \hat{\mathbf{u}}_3=[12.75\ 8.24\ 2.62]^T$	$\hat{\pi}_3=0.283,\ \hat{\mathbf{u}}_3=[12.32\ 7.63\ 2.19\ 5.75]^T$ $\hat{\pi}_5=0.527,\ \hat{\mathbf{u}}_5=[4.57\ 13.43\ 3.53\ 9.99]^T$ $\hat{\pi}_6=0.190,\ \hat{\mathbf{u}}_6=[2.83\ 4.85\ 11.97\ 6.28]^T$
$N=500$	$\hat{\pi}_1=0.346,\ \hat{\mathbf{u}}_1=[10.26\ 5.95\ 2.13]^T$ $\hat{\pi}_3=0.654,\ \hat{\mathbf{u}}_3=[3.94\ 12.11\ 2.82]^T$	$\hat{\pi}_2=0.505,\ \hat{\mathbf{u}}_2=[4.28\ 12.12\ 3.06\ 9.35]^T$ $\hat{\pi}_4=0.200,\ \hat{\mathbf{u}}_4=[3.09\ 4.73\ 11.74\ 5.69]^T$ $\hat{\pi}_6=0.295,\ \hat{\mathbf{u}}_5=[12.14\ 7.04\ 2.44\ 5.71]^T$
$N=1000$	$\hat{\pi}_3=0.645,\ \hat{\mathbf{u}}_3=[3.99\ 11.75\ 3.09]^T$ $\hat{\pi}_4=0.355,\ \hat{\mathbf{u}}_4=[10.09\ 6.09\ 2.07]^T$	$\hat{\pi}_3=0.498,\ \hat{\mathbf{u}}_3=[9.98\ 6.00\ 2.06\ 5.11]^T$ $\hat{\pi}_4=0.200,\ \hat{\mathbf{u}}_4=[3.15\ 5.07\ 12.21\ 6.18]^T$ $\hat{\pi}_6=0.302,\ \hat{\mathbf{u}}_6=[4.21\ 12.55\ 3.29\ 9.06]^T$
	Model C: $\pi_1=0.35,\ \mathbf{u}_1=[3\ 5\ 12\ 6\ 8]^T$ $\pi_2=0.3,\ \mathbf{u}_2=[4\ 12\ 3\ 9\ 2]^T$ $\pi_3=0.25,\ \mathbf{u}_3=[10\ 6\ 2\ 5\ 20]^T$ $\pi_4=0.1,\ \mathbf{u}_4=[3\ 12\ 5\ 6\ 16]^T$	**Model D**: $\pi_1=0.3,\ \mathbf{u}_1=[3\ 5\ 12\ 6\ 8\ 20]^T$ $\pi_2=0.25,\ \mathbf{u}_2=[4\ 12\ 3\ 9\ 2\ 18]^T$ $\pi_3=0.2,\ \mathbf{u}_3=[10\ 6\ 2\ 5\ 20\ 30]^T$ $\pi_4=0.15,\ \mathbf{u}_4=[3\ 12\ 5\ 6\ 16\ 4]^T$ $\pi_5=0.1,\ \mathbf{u}_5=[19\ 6\ 18\ 20\ 14\ 3]^T$
$N=100$	$\hat{\pi}_1=0.333,\ \hat{\mathbf{u}}_1=[3.21\ 5.64\ 12.91\ 6.61\ 8.19]^T$ $\hat{\pi}_5=0.111,\ \hat{\mathbf{u}}_5=[4.83\ 20.38\ 7.25\ 8.44\ 24.59]^T$ $\hat{\pi}_6=0.296,\ \hat{\mathbf{u}}_6=[5.28\ 15.25\ 4.16\ 12.47\ 2.45]^T$ $\hat{\pi}_8=0.260,\ \hat{\mathbf{u}}_8=[13.86\ 8.91\ 2.89\ 8.07\ 28.74]^T$	$\hat{\pi}_2=0.103,\ \hat{\mathbf{u}}_2=[26.33\ 9.10\ 23.00\ 27.72\ 17.39\ 3.16]^T$ $\hat{\pi}_3=0.197,\ \hat{\mathbf{u}}_3=[12.46\ 7.24\ 2.58\ 5.63\ 25.23\ 38.30]^T$ $\hat{\pi}_5=0.150,\ \hat{\mathbf{u}}_5=[2.77\ 14.00\ 5.94\ 6.72\ 16.52\ 4.86]^T$ $\hat{\pi}_7=0.237,\ \hat{\mathbf{u}}_7=[4.59\ 16.85\ 3.72\ 11.63\ 2.43\ 24.29]^T$ $\hat{\pi}_8=0.313,\ \hat{\mathbf{u}}_8=[3.10\ 4.89\ 10.76\ 5.61\ 7.37\ 18.20]^T$
$N=200$	$\hat{\pi}_3=0.352,\ \hat{\mathbf{u}}_3=[2.86\ 4.72\ 10.83\ 4.94\ 8.05]^T$ $\hat{\pi}_6=0.088,\ \hat{\mathbf{u}}_6=[2.81\ 13.44\ 4.05\ 7.14\ 19.07]^T$ $\hat{\pi}_7=0.248,\ \hat{\mathbf{u}}_7=[9.71\ 6.25\ 1.99\ 5.09\ 20.92]^T$ $\hat{\pi}_8=0.312,\ \hat{\mathbf{u}}_8=[4.52\ 12.88\ 3.09\ 9.43\ 2.24]^T$	$\hat{\pi}_1=0.201,\ \hat{\mathbf{u}}_1=[10.79\ 6.67\ 2.02\ 5.68\ 23.60\ 32.70]^T$ $\hat{\pi}_7=0.294,\ \hat{\mathbf{u}}_7=[3.46\ 5.25\ 12.31\ 6.64\ 7.85\ 22.96]^T$ $\hat{\pi}_8=0.253,\ \hat{\mathbf{u}}_8=[4.68\ 14.87\ 3.40\ 10.83\ 2.48\ 20.42]^T$ $\hat{\pi}_9=0.103,\ \hat{\mathbf{u}}_9=[23.04\ 7.42\ 21.75\ 24.69\ 14.51\ 3.12]^T$ $\hat{\pi}_{10}=0.149,\ \hat{\mathbf{u}}_{10}=[3.29\ 13.72\ 4.65\ 7.08\ 17.48\ 3.73]^T$
$N=500$	$\hat{\pi}_1=0.107,\ \hat{\mathbf{u}}_1=[3.15\ 12.11\ 5.43\ 6.03\ 17.21]^T$ $\hat{\pi}_4=0.307,\ \hat{\mathbf{u}}_4=[4.39\ 12.62\ 3.23\ 9.49\ 2.10]^T$ $\hat{\pi}_7=0.337,\ \hat{\mathbf{u}}_7=[3.03\ 5.14\ 13.19\ 6.82\ 8.81]^T$ $\hat{\pi}_8=0.249,\ \hat{\mathbf{u}}_8=[11.15\ 6.83\ 2.23\ 5.59\ 23.19]^T$	$\hat{\pi}_1=0.201,\ \hat{\mathbf{u}}_1=[9.36\ 5.94\ 2.04\ 5.03\ 19.54\ 30.52]^T$ $\hat{\pi}_4=0.100,\ \hat{\mathbf{u}}_4=[19.78\ 6.03\ 18.32\ 20.87\ 14.66\ 3.36]^T$ $\hat{\pi}_5=0.152,\ \hat{\mathbf{u}}_5=[2.82\ 10.35\ 4.31\ 5.25\ 14.73\ 3.79]^T$ $\hat{\pi}_8=0.246,\ \hat{\mathbf{u}}_8=[4.22\ 12.28\ 3.22\ 9.64\ 2.07\ 17.96]^T$ $\hat{\pi}_{10}=0.301,\ \hat{\mathbf{u}}_{10}=[3.57\ 6.49\ 15.79\ 7.34\ 9.70\ 24.82]^T$
$N=1000$	$\hat{\pi}_2=0.300,\ \hat{\mathbf{u}}_2=[3.88\ 11.89\ 3.06\ 9.49\ 1.99]^T$ $\hat{\pi}_3=0.343,\ \hat{\mathbf{u}}_3=[3.22\ 4.97\ 12.24\ 6.18\ 8.23]^T$ $\hat{\pi}_4=0.253,\ \hat{\mathbf{u}}_4=[9.06\ 5.38\ 1.83\ 4.72\ 17.93]^T$ $\hat{\pi}_8=0.104,\ \hat{\mathbf{u}}_8=[2.83\ 13.13\ 5.53\ 6.00\ 16.88]^T$	$\hat{\pi}_1=0.150,\ \hat{\mathbf{u}}_1=[3.02\ 12.71\ 5.08\ 6.54\ 17.97\ 4.23]^T$ $\hat{\pi}_3=0.251,\ \hat{\mathbf{u}}_3=[3.93\ 11.75\ 2.88\ 8.77\ 1.98\ 18.17]^T$ $\hat{\pi}_8=0.101,\ \hat{\mathbf{u}}_8=[18.69\ 5.53\ 17.26\ 19.74\ 13.64\ 3.03]^T$ $\hat{\pi}_9=0.298,\ \hat{\mathbf{u}}_9=[3.12\ 5.12\ 12.47\ 6.50\ 8.35\ 20.64]^T$ $\hat{\pi}_{10}=0.200,\ \hat{\mathbf{u}}_{10}=[9.42\ 5.85\ 1.94\ 4.57\ 19.31\ 28.43]^T$

Table 2
Estimation of skewed data's distribution with the proposed VI-based method. Since the two distributions are well separated from each other, the estimated mixture weights are $\hat{\pi}_1=0.3$ and $\hat{\pi}_2=0.7$ for all the cases. Thus, the estimated mixture weights are not listed in this table.

Model: $\pi_1=0.3,\ \mathbf{u}_1=[2\ 1000\ 2\ 1000]^T$
$\pi_2=0.7,\ \mathbf{u}_2=[1000\ 2\ 1000\ 2]^T$

Number of samples	$N=100$	$N=200$
Estimated parameters	$\mathbf{u}_1=[1.73\ 1003.91\ 1.82\ 1005.28]^T$ $\mathbf{u}_2=[949.43\ 1.93\ 955.83\ 1.95]^T$	$\mathbf{u}_1=[2.53\ 1105.41\ 2.07\ 1162.83]^T$ $\mathbf{u}_2=[1105.41\ 2.03\ 1109.72\ 1.92]^T$
Number of samples	$N=500$	$N=1000$
Estimated parameters	$\mathbf{u}_1=[2.17\ 1085.25\ 2.05\ 1084.29]^T$ $\mathbf{u}_2=[1023.20\ 2.21\ 1022.11\ 2.12]^T$	$\mathbf{u}_1=[2.19\ 1085.28\ 2.28\ 1084.16]^T$ $\mathbf{u}_2=[1023.30\ 2.04\ 1021.92\ 2.08]^T$

Table 3
Comparisons of the KL divergences (39) obtained by the using proposed VI-based method and the ML estimation based method [25]. Smaller KL divergence indicates a better performance.

Parameters	Method	$N=10$	$N=50$	$N=100$
$\mathbf{u}=[3\ 5\ 8]^{\mathrm{T}}$	VI	1.18×10^{-7}	4.75×10^{-9}	9.58×10^{-10}
	ML	1.79×10^{-6}	5.64×10^{-7}	2.43×10^{-7}
$\mathbf{u}=[10\ 6\ 20]^{\mathrm{T}}$	VI	1.09×10^{-7}	2.85×10^{-9}	3.95×10^{-10}
	ML	6.92×10^{-6}	6.84×10^{-6}	6.81×10^{-6}

that the Bayesian estimation provides a reliable estimate, especially when the amount of observed data is small. To illustrate the advantages of the proposed VI-based method over the ML estimation [25],[5] we evaluated the Kullback–Leibler (KL) divergence from the true predictive distribution to the estimated one as

$$\mathrm{KL}(f(\mathbf{x}|\mathbf{X})\|f(x;\widehat{\mathbf{u}}))=\int f(\mathbf{x}|\mathbf{X})\log\frac{f(\mathbf{x}|\mathbf{X})}{f(x;\widehat{\mathbf{u}})}\,d\mathbf{x}, \tag{39}$$

where the predictive distribution was calculated by the importance sampling method [3] as directly sampling from the posterior distribution $f(\mathbf{u}|\mathbf{X})$ is not feasible. The smaller the KL divergence is, the better the performance is. Here, the estimated parameters are either the point estimates (posterior means) from the proposed VI-based method or the point estimates from the ML estimation based method. The comparison results are shown in Table 3. Both estimation methods yield accurate estimates (very small KL divergence), especially when the amount of observation is relatively large. Moreover, the VI-based method leads to smaller KL divergences than the ML estimation, for different amounts of observed data. We believe that this is due to the advantage of Bayesian estimation. 50 rounds of simulations were evaluated for each N and the mean values are reported.

Moreover, we also compare the numerical complexities of the proposed VI-based method with the ML estimation method in [25]. When stopping criterion is reached, it is observed that (see Fig. 2) the proposed VI-based method converges after 70–120 iterations, with the synthesized "normal" data (the case with the highly skewed data requires more iterations). The ML estimation algorithm introduced in [25] also requires the same rounds of simulations. The main calculation costs in both methods are from matrix addition and multiplication. However, different from the VI-based algorithm, the ML estimation involves an extra numerical gradient method (*e.g.*, Newton–Raphson algorithm) in the calculation of the maximization step. Therefore, the ML estimation for DMM is more computationally costly than the VI-based method for DMM. To verify this, we recorded the computational time required by different methods. The average run time (with 100 iterations for both methods), obtained from 50 rounds of simulations with $N=10\,000$ samples, is shown in Table 4. It can be observed that the VI-based method for DMM has less overall computational cost than the ML estimation for DMM.

4.1.4. Comparisons with other Bayesian estimation method

The VI-based method proposed in this paper uses the SLB approximation while the method proposed in [12] used the MLB approximation. Although both methods carry out the Bayesian estimation of DMM under the EFA framework, different objective functions are maximized during each iteration. It has already been discussed in Section 3.6.2 that only the SLB based method can guarantee the convergence theoretically. In this section, we compare these two methods quantitatively. With a known DMM, 10 000 samples were generated. The above-mentioned two algorithms were applied to estimate the DMM. After convergence, we calculated the variational objective function

$$\mathcal{L}=\mathbb{E}_{\mathcal{Z}}[\ln f(\mathbf{X},\mathcal{Z})]-\mathbb{E}_{\mathcal{Z}}[f(\mathcal{Z})], \tag{40}$$

which is actually the true lower-bound to the model evidence in the VI framework, to examine which approximation is better. With the obtained posterior distribution $f^*(\mathbf{Z})$, the variational objective function $\mathcal{L}$ can be calculated numerically by sampling from the obtained posterior distribution $f^*(\mathbf{Z})$. Hence, we got two values, $\mathcal{L}_{\mathrm{SLB}}$ and $\mathcal{L}_{\mathrm{MLB}}$, from two algorithms. As expected, both the SLB and the MLB based method converged and estimated the parameters and the model complexity (in terms of the number of mixture components) properly. The lower-bounds estimated by these two methods are listed Table 5. The SLB approximation based method yields higher lower-bound value than the MLB approximation based method in [12]. This fact indicates that the SLB approximation is tighter than the MLB approximation. To further check the stability of these two methods, we draw box plots for both cases. Fig. 4 shows the comparisons between these two methods. It can be observed that the SLB approximation based method has compact lower-bounds range, higher median, and less outliers. Therefore, we can conclude that the SLB approximation based method is superior to the MLB approximation based one.

Table 4
Comparisons of the overall computational cost with 10 000 samples generated from each model (measured in seconds on a modern Dell Precision Workstation).

Method	**Model A**	**Model B**	**Model C**	**Model D**
VI	3.52	5.58	8.21	9.57
ML	5.63	8.72	13.14	16.34

4.2. Real data evaluation

Compared to the conventional ML estimation, the Bayesian estimation conveys a robust modeling performance. It can yield not only a reasonable parameter estimate but also the complexity of the model. In real data evaluation, we study two important multimedia signal processing applications: (1) the LSF parameter quantization in speech coding and (2) multiview depth image enhancement in free-view point television. The latter one is also an unsupervised learning problem.

4.2.1. LSF parameters quantization

In speech coding, quantization of the LSF parameters plays an essential role [33]. The LSF parameters, which are widely used in speech coding, have a number of properties: the support range is bounded, the elements in a LSF vector are ordered, and the filter stability can be easily checked [33]. Such properties have been studied and considered explicitly when implementing an efficient LSF vector quantization (VQ) [34]. The PDF-optimized VQ has been previously shown to be more efficient than the VQ based only on training data [35,36]. In our previous work [13], we proposed a new PDF-optimized VQ method based on the Dirichlet distribution to quantize the LSF parameters. The proposed method, which is based on the ML estimation and carried out by the EM algorithm, showed an improvement over the conventional GMM based method.

[5] The EM algorithm was firstly (to our best knowledge) proposed in [10]. However, the proposed method in [10] estimated the logarithm of the parameter u_k, instead of the parameter itself. In this paper, we take the method suggested in [25], in which the parameter u_k was estimated directly. The ML estimation method proposed in [25] is for a DMM. It can also be applied to estimate a single Dirichlet distribution.

Table 5
Comparisons of the proposed VI-based method (*via* SLB approximation) with the Bayesian estimation method (*via* MLB approximation) in [12]. The simulations were run 50 rounds with 10 000 generated samples and the averages are reported. The parameters settings in model A and B are the same as those in Table 5.

Model A	$\mathcal{L}_{\mathrm{SLB}}-\mathcal{L}_{\mathrm{MLB}}$	Model B	$\mathcal{L}_{\mathrm{SLB}}-\mathcal{L}_{\mathrm{MLB}}$
$\pi_1=0.65,\ \mathbf{u}_1=[4\ 12\ 3]^{\mathrm{T}}$ $\pi_2=0.35,\ \mathbf{u}_2=[10\ 6\ 2]^{\mathrm{T}}$	1.51×10^{-2}	$\pi_1=0.2,\ \mathbf{u}_1=[3\ 5\ 12\ 6]^{\mathrm{T}}$ $\pi_2=0.5,\ \mathbf{u}_2=[4\ 12\ 3\ 9]^{\mathrm{T}}$ $\pi_3=0.3,\ \mathbf{u}_3=[10\ 6\ 2\ 5]^{\mathrm{T}}$	0.71×10^{-2}

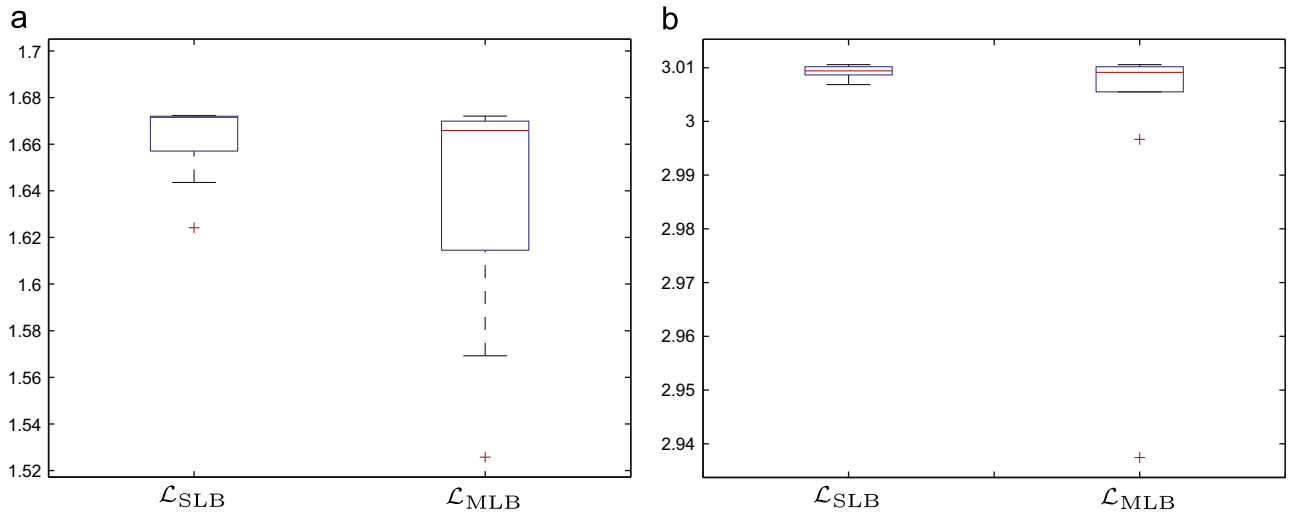

Fig. 4. Comparisons of the $\mathcal{L}_{\mathrm{SLB}}$ and $\mathcal{L}_{\mathrm{MLB}}$ values *via* box plot. The central mark is the median, the edges of the box are the 25th and 75th percentiles. The outliers are marked individually. For each model, the simulations were run 20 rounds with 10 000 generated samples. The parameters in models A (a) and B (b) are the same as those in Table 5.

For a K-dimensional LSF vector $\mathbf{s}=[s_1,s_2,\ldots,s_K]^{\mathrm{T}}$, the elements are bounded in the interval $(0,\pi)$ and are strictly ordered as

$$0<s_1<s_2<\cdots<s_K<\pi. \tag{41}$$

To explicitly exploit such properties, we firstly calculate the intervals between two adjacent LSF elements (including 0 and π) as

$$\mathbf{x}=[x_1,x_2,\ldots,x_{K+1}]^{\mathrm{T}}=\frac{1}{\pi}[s_1,s_2-s_1,\ldots,s_K-s_{K-1},\pi-s_K]^{\mathrm{T}}. \tag{42}$$

The obtained vector $\mathbf{x}$ is referred to as the ΔLSF vector. Since all the elements in $\mathbf{x}$ are positive and the summation of all the elements equals one, we model the underlying distribution of $\mathbf{x}$ by a DMM.

By assuming a linear predictive model of the speech signal, the 16-dimensional linear predictive coding (LPC) vector parameters were extracted from the wide-band TIMIT [37] database. Fig. 5 shows the marginal histograms for the elements in the extracted LSF vector, where the boundary and ordering properties are clearly illustrated. Based on the above procedure, the LSF parameters were transformed to the ΔLSF representation and modeled by a DMM [13]. In this paper, we applied the Bayesian DMM proposed above to estimate the model. With the bit allocation, decorrelation, and quantization strategies in [13], the BDMM based VQ (BDVQ) was designed. Meanwhile, we also carried out a Bayesian GMM [3] based VQ (BGVQ) as the benchmark, since the GMM based VQ is the state-of-the-art PDF-optimized VQ [35].

The distortions between the true and quantized LSF parameters, which are measured in mean squared error (MSE) and log spectral distortion (LSD), were utilized to evaluate the performance of the Bayesian DMM based VQ (BDVQ). The MSE represents the average Euclidean distance between the true LSF vector and the quantized one, the smaller the better. The LSD is more correlated with the speech perceptual distortion than the MSE,

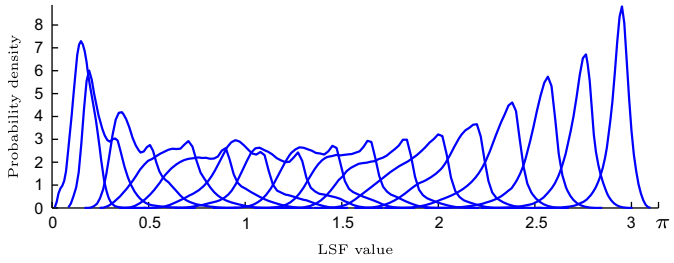

Fig. 5. Illustration of the bounded and ordering properties of the LSF parameters. The plots from left to right are corresponding to the histograms from the 1st to the 16th dimension [13].

thus it is widely used as the objective measure of the LSF quantizer [38]. The evaluation criterion for the LSD is the bit rate that the quantizer reaches the "transparent coding", which is

1. 1 dB LSD on average,
2. less than 2% outliers in the 2–4 dB range,
3. no outlier larger than 4 dB.

We also evaluated the EM-DMM based VQ (EMDVQ) [13] and the EM-GMM based VQ (EMGVQ) [36] to make a comprehensive comparison. In the training procedure, all the four models, the Bayesian DMM, the EM DMM, the Bayesian GMM, and the EM GMM, were initialized with 32 mixture components. The estimation carried out by the Bayesian procedure can find the optimal number of mixture components by itself. For the EM algorithm based method, we used the Bayesian information criterion (BIC) [39] to decide the model complexity. On average, both the Bayesian DMM and the EM DMM kept around 7 active mixture components/125 free parameters while the Bayesian GMM and the EM GMM kept 10 mixture components/329 free parameters. Thus,

the BDVQ has a less model complexity (in terms of the number of mixture components and free parameters), which is favorable for practical speech coding application. The comparison results are listed in Table 6. The comparisons show that the BDVQ performs the best among all the four methods, which gives the lowest bit rate and the smallest MSE. Compared to the EMDVQ, BGVQ, and the EMGVQ, the BDVQ outperforms by about 1, 3, and 4 bits/vector for all the criteria, respectively. The reported values are the mean of 50 rounds of simulations.

4.2.2. Multiview depth image enhancement

Free-viewpoint television (FTV) will enable viewers to experience a dynamic natural 3D-depth impression while freely choosing their viewpoint of real world scenes [26]. This has been made possible by recent advances in autostereoscopic multiview display technology which permits viewing of scenes from a range of perspectives for multiple viewer [41]. Multiview displays require a large number of views at the receiver side to have a seamless transition among interactively selected stereo pairs [42]. This requires to capture, store, and transmit an enormous amount of multiview video (MVV) [43]. In recent years, many compression techniques have been proposed for MVV imagery [43,44]. However, the resulting transmission cost is approximately proportional to the number of coded views. Therefore, a large number of views cannot be efficiently transmitted using existing techniques. The transmission efficiency can be improved significantly by utilizing depth maps [42]. A depth map is a single channel gray scale image where each pixel represents the shortest distance between the corresponding object point in the natural scene and the given camera plane. Given a small subset of MVV imagery and its corresponding set of multiview depth (MVD) images, an arbitrary number of views can be synthesized by using depth image based rendering (DIBR) [45]. Usually, depth maps are obtained by establishing stereo correspondences between two or more camera images at different viewpoints by a matching criterion [46]. The accuracy of the stereo matching affects the resulting depth estimates. Despite a number of optimization techniques that are used to refine depth estimates [47,48], the resulting depth maps at different viewpoints usually lack inter-view consistency. However, this inconsistency affects the quality of view interpolation negatively [49]. Inter-view inconsistencies usually arise from complexity-constrained local estimation techniques as depicted in Fig. 6 for the 1D-parallel camera array setting. In this camera setting, all optical centers of the cameras are parallel to each other and all rotation matrices are identical.

Table 6
Comparisons of transparent coding performances.

Method	bits/vec.	MSE ($\times 10^{-3}$)	LSD (dB)	LSD outliers (%)	
				2–4 dB	>4 dB
BDVQ	51	2.5	1.0087	1.434	0.000
	52	1.8	0.8890	1.271	0.000
EMDVQ	52	2.6	1.0239	1.802	0.000
	53	2.4	0.9904	1.523	0.000
BGVQ	54	2.9	1.0160	1.612	0.035
	55	2.7	0.9834	1.259	0.035
EMGVQ	55	3.2	1.0035	1.180	0.038
	56	2.5	0.9576	0.943	0.032

In previous work by some of the authors [50], a general model-based framework for depth map enhancement has been proposed. It exploits the conditional dependency between color and depth. The framework uses the view imagery and the corresponding depth maps at their respective viewpoints and performs a RGB color pixel classification in the view imagery by producing a generative model based on a mixture of Gaussian distributions. The model parameters are estimated by variational Bayesian inference [3]. To make the procedure insensitive to the absolute luminance, we transform the MVV imagery from the RGB space to the XYZ color space [51]. This color space has virtual primaries and all spectral matching curves are positive. The chromaticity values are obtained by normalization such that they sum to one, that is

$$x=\frac{X}{X+Y+Z},\quad y=\frac{Y}{X+Y+Z},\quad z=\frac{Z}{X+Y+Z}. \tag{43}$$

With $\mathbf{v}=[x\ y\ z]^{\mathrm{T}}$, each pixel in the color image is represented by that vector $\mathbf{v}$ which has nonnegative elements that sum to one.

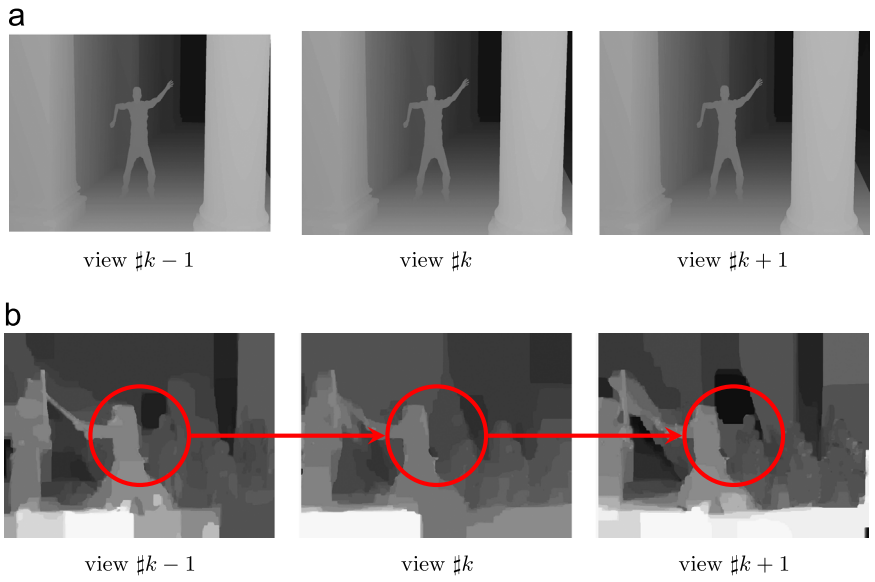

Fig. 6. An example of inter-view inconsistency among multiview depth maps at different viewpoints for multiview video imagery as provided by [40]. For 1D-parallel camera arrays, the depth value of a unique 3D point is the same in all depth maps, but located at different positions in the maps. Therefore, depth observations at different viewpoints should be consistent and related areas in different viewpoints should show the same depth values, but shifted. The Dancer test data in (a) is a synthetic test material and has consistent depth maps across all viewpoints. This is not the case in (b) for the estimated Kendo depth imagery, where red circles mark prominent inconsistent areas in the depth maps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

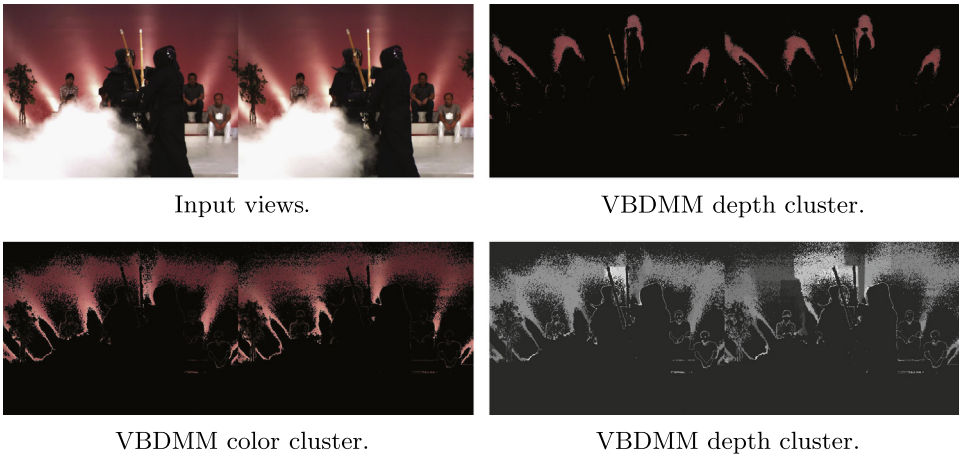

Fig. 7. An example of color and corresponding depth classification results for Kendo. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Table 7
Objective quality of the synthesized virtual views.

Test sequence	Input views	Virtual view	VSRS 3.5 (dB)				
			MPEG depth (a)	RGB VBGMM + K-means depth (b)	RGB VBGMM + mean-shift depth (c)	XYZ VBGMM + mean-shift depth (d)	xyz VBDMM + mean-shift depth (e)
Kendo	3, 5	4	36.54	36.72	36.83	36.88	39.35
Lovebird1	6, 8	7	28.50	28.67	28.81	28.85	29.01

Such properties fit the Dirichlet distribution. Thus, by assuming that the pixel values are statistically independent from each other, we model the underlying distribution of the pixel color (in the chromaticity space) by a DMM. In this modeling procedure, the pixels that have a similar color should be clustered to the same mixture component. In general, the number of color clusters in an image is not known in advance. Hence, the proposed Bayesian DMM algorithm can be applied to estimate the best number of mixture components (*i.e.*, the number of color clusters) for each image. This is actually an unsupervised learning problem. After clustering the pixels, the segmented color images can be used to further enhance the associated depth maps by the mean-shift algorithm [32]. The members of a specific color cluster have similar colors, while on the contrary, the members of a specific depth cluster may have distinct depth values. As foreground and background object points can have similar colors, foreground object points have different depth values compared to background object points. An object point with a given color which is visible from all viewpoints should have the same depth value in all depth maps. However, such points usually have different depth values in the cluster due to inter-view inconsistencies. The mean-shift algorithm will help us to resolve this ambiguity by further sub-clustering depth values and by assigning mean depth values to sub-clusters. Fig. 7 shows color clusters and the associated depth clusters for concatenated color images and depth maps, respectively. More details can be found in [32].

FTV viewers will enjoy both real camera captured views at given camera locations and synthesized virtual views at other arbitrary viewpoints. Therefore, the proposed Bayesian DMM-based algorithm is evaluated in two steps. First, the depth imagery at two viewpoints is improved. Second, a virtual view for a given viewpoint is synthesized by employing the MPEG view synthesis reference software (VSRS) 3.5 [52]. VSRS is a DIBR algorithm that interpolates a virtual view at an arbitrary intermediate viewpoint. For this, it uses two reference views, left and right, the two corresponding reference depth maps, and the corresponding camera parameters [45,53]. The camera parameters include the 3×3 intrinsic camera parameter matrix, the 3×3 camera rotation matrix, and the 3×1 camera translation vector. These camera parameters define the camera projection matrix [53]. Using a basic pinhole camera model, DIBR forms an image by projecting reference image pixels to 3D world points by using the corresponding depth information and the reference camera projection matrix. The resulting 3D word points are projected back to the image plane at the virtual viewpoint by using the virtual camera projection matrix. Usually, this process is known as 3D image warping [54–56]. Some areas in the warped views at the virtual viewpoint may have holes because the warping process is affected by disocclusion and quantized depth values. Disocclusions are detected by checking the discontinuities in the depth map at the virtual viewpoint. The missing information in one warped view is likely to be available in the warped version of another reference view. Therefore, information from other warped views is used to fill the hole areas created by disocclusion [45]. The warped views at the virtual viewpoint are blended to generate the final virtual view. If some holes still remain in this view, their intensity values are filled by using other techniques such as inpainting [57]. In the following experiments, VSRS operates in the 1D-parallel synthesis mode at half-pel precision and uses our consistent depth maps [58]. Finally, we measure the objective quality of the synthesized views in terms of the peak signal-to-noise ratio (PSNR) with respect to a captured view of a real camera at the same viewpoint. For our experiments, we use standard MVV test data as provided by MPEG [40].

Table 7 shows a comparison of the luminance signal Y-PSNR (in dB) of the virtual views as synthesized by VSRS 3.5 with the help of (a) MPEG depth maps, (b) enhanced depth maps from [50], (c) enhanced depth maps using VBGMM in RGB space and mean shift, (d) enhanced depth maps using VBGMM in XYZ space and mean shift, and (e) enhanced depth maps from the proposed Bayesian DMM-based approach. The presented depth enhancement algorithm offers improvements in image quality in the range from 0.5 dB to 2.8 dB, when compared to image qualities

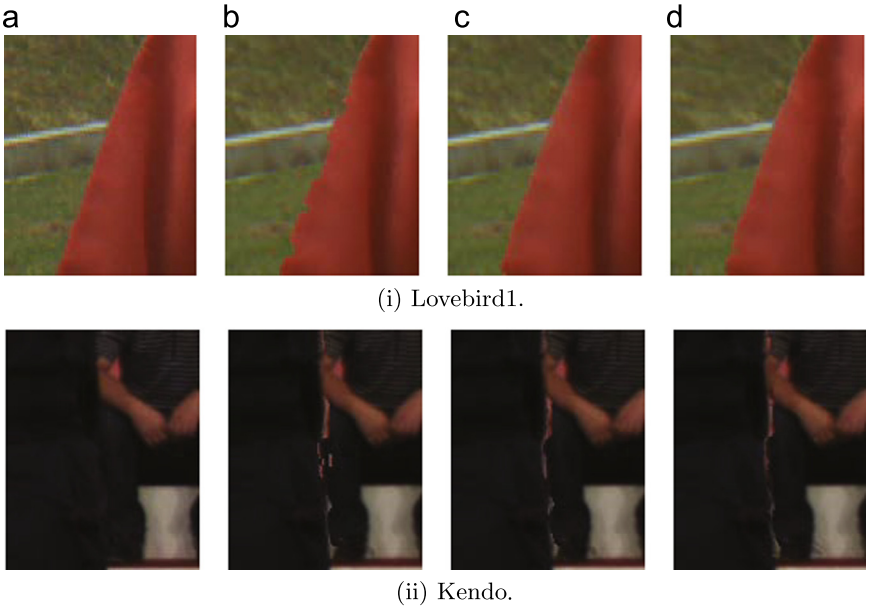

Fig. 8. Selected areas of synthesized virtual views of the test sequences as generated by VSRS 3.5 using (b) MPEG depth maps, (c) improved depth maps by [50], and (d) enhanced depth maps from the proposed VBDMM+Mean-shift based algorithm. Synthesized virtual views at full resolution are available at http://people.kth.se/~prara/research/dmmresults.zip.

as obtained by standard MPEG depth maps. The improvement in quality highly depends on the input reference depth maps at various viewpoints. In Table 7, the best results are presented. Fig. 8 shows the efficiency of the proposed depth map enhancement algorithm. Note the improved visual quality of the virtual views when compared with MPEG depth maps. Specially, artifacts around the edges in the synthesized views have been significantly reduced. The improvement in the virtual view quality of Kendo is due to better color classification results by using VBDMM. Hence, this is a promising algorithm for improving the visual quality of FTV.

Besides improving the quality of FTV, our DMM-based approach offers less model complexity than the GMM-based approach. As the Bayesian DMM-based method yields better visual quality than the Bayesian GMM-based method with even less model complexity, it is preferable in practice.

5. Conclusion

The Bayesian estimation of a statistical model is, in general, preferable to the maximum likelihood (ML) estimation. To avoid the numerical calculation in the maximum likelihood estimation of the parameters in a Dirichlet mixture model (DMM), we proposed a novel Bayesian estimation method based on the variational inference framework. The main contribution of this paper is to derive an analytically tractable solution for approximating the posterior distribution of the parameters, by utilizing the relative convex properties of the multivariate log-inverse-beta functions. Since the single lower-bound to the variational objective function is derived and maximized during each iteration, the proposed algorithm is guaranteed to converge. Compared to the ML estimation, the proposed Bayesian estimation method can prevent the numerical search, estimate the number of mixture components automatically and accurately, and overcome the over-fitting problem.

With synthesized data validation, the efficiency and accuracy of the proposed algorithm were verified. For the real data evaluation, we demonstrated the performance of the proposed Bayesian DMM method with two important multimedia signal processing applications. In the practical line spectral frequency (LSF) parameter quantization, the proposed Bayesian estimation method showed better transparent coding performance, compared to the ML estimation based method and the state-of-the-art Gaussian mixture model (GMM) based method. For the purpose of improving the free-viewpoint TV quality, the Bayesian DMM based method yielded a significant improvement in the depth map enhancement and outperformed the MPEG standard and the recently proposed Bayesian GMM based method.

Conflict of interest statement

None declared.

Appendix A. Proof of Theorem 1

The MLIB function can be factorized as

$$\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})}=\ln\frac{\Gamma(u_{ki}+\sum_{m\neq k}u_{mi})}{\Gamma(u_{ki})\Gamma(\sum_{m\neq k}u_{mi})}+\ln\frac{\Gamma(\sum_{m\neq k}u_{mi})}{\prod_{m\neq k}\Gamma(u_{mi})}. \tag{A.1}$$

The first term in the above equation is the log-inverse-beta (LIB) function $\ln\Gamma(x+y)/\Gamma(x)\Gamma(y)$. As proposed and proven in [8, Properties 3.3 and 3.4], the LIB function is convex relative to $\ln x$ when $y>1$. Thus, the MLIB function is convex in $\ln u_{ki}$ (relative [29] in u_{ki}) when $\sum_{m\neq k}u_{mi}>1$. Its first-order Taylor expansion in terms of $\ln u_{ki}$ around *any* point is a lower-bound to it.

By considering u_{1i} as the only variable, the MLIB function can be approximated by its first-order Taylor expansion in terms of $\ln u_{1i}$ around $\ln\overline{u}_{1i}$ as

$$\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})}\geq\ln\frac{\Gamma(\overline{u}_{1i}+\sum_{k=2}^{K+1}u_{ki})}{\Gamma(\overline{u}_{1i})\prod_{k=2}^{K+1}\Gamma(u_{ki})}+\left[\psi\left(\overline{u}_{1i}+\sum_{k=2}^{K+1}u_{ki}\right)-\psi(\overline{u}_{1i})\right]\overline{u}_{1i}(\ln u_{1i}-\ln\overline{u}_{1i}), \tag{A.2}$$

where $\overline{u}$ is the expected value of u and $\psi(\cdot)$ is the digamma function defined as $\psi(x)=\partial\ln\Gamma(x)/\partial x$.

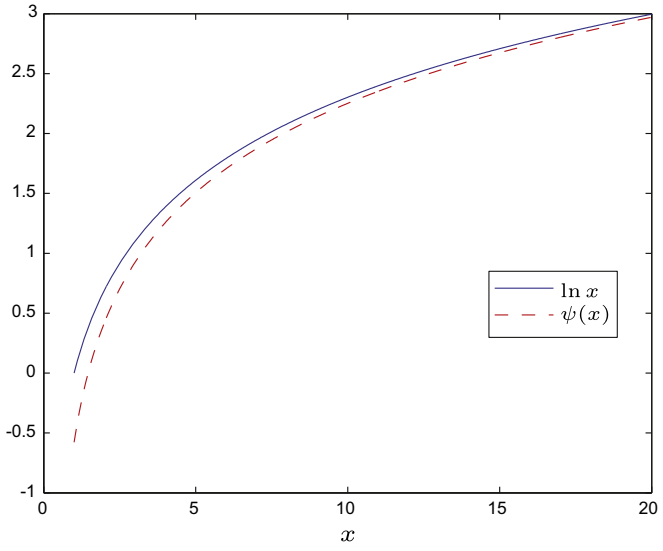

Fig. B1. Comparison of $\ln x$ and $\psi(x)$.

Next, if we consider u_{2i} as the variable, the MLIB function can be further approximated as

$$\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})} \geq \ln\frac{\Gamma(\overline{u}_{1i}+\overline{u}_{2i}+\sum_{k=3}^{K+1}u_{ki})}{\Gamma(\overline{u}_{1i})\Gamma(\overline{u}_{2i})\prod_{k=3}^{K+1}\Gamma(u_{ki})} + \left[\psi\left(\overline{u}_{1i}+\overline{u}_{2i}+\sum_{k=3}^{K+1}u_{ki}\right)-\psi(\overline{u}_{2i})\right]\overline{u}_{2i}(\ln u_{2i}-\ln\overline{u}_{2i}) + \left[\psi\left(\overline{u}_{1i}+\sum_{k=2}^{K+1}u_{ki}\right)-\psi(\overline{u}_{1i})\right]\overline{u}_{1i}(\ln u_{1i}-\ln\overline{u}_{1i}). \quad (A.3)$$

By repeating the above procedure for all the remaining variables, we can obtain a lower-bound approximation to the MLIB function, in terms of all the variables, as

$$\ln\frac{\Gamma(\sum_{k=1}^{K+1}u_{ki})}{\prod_{k=1}^{K+1}\Gamma(u_{ki})} \geq \ln\frac{\Gamma(\sum_{k=1}^{K+1}\overline{u}_{ki})}{\prod_{k=1}^{K+1}\Gamma(\overline{u}_{ki})} + \sum_{k=1}^{K+1}\left[\psi\left(\sum_{m=1}^{k}\overline{u}_{mi}+\sum_{l=k+1}^{K+1}u_{li}\right)-\psi(\overline{u}_{ki})\right]\overline{u}_{ki}(\ln u_{ki}-\ln\overline{u}_{ki}). \quad (A.4)$$

Appendix B. Tightening the lower-bound

As several approximations were used in the derivations of the lower-bound to the expectation of the MLIB function, it is interesting to discuss the tightness of this lower-bound.

Firstly, let us look at the first-order Taylor expansion of the LIB function $\ln\Gamma(x+y)/\Gamma(x)\Gamma(y)$ in terms of $\ln x$ around *any* point $\widetilde{\ln x}$. For the expectation of the LIB function, we have the following inequality:

$$\mathbb{E}_{f(x)}\left[\ln\frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}\right] \geq \mathbb{E}_{f(x)}\left\{\ln\frac{\Gamma(e^{\widetilde{\ln x}}+y)}{\Gamma(e^{\widetilde{\ln x}})\Gamma(y)}+[\psi(e^{\widetilde{\ln x}}+y) -\psi(e^{\widetilde{\ln x}})]e^{\widetilde{\ln x}}(\ln x-\widetilde{\ln x})\right\}. \quad (B.1)$$

Taking the derivative of the RHS of (B.1) with respect to $\widetilde{\ln x}$ can maximize this first-order Taylor expansion. With some calculations, the optimal $\widetilde{\ln x}$ is

$$\widetilde{\ln x}^{*} = \mathbb{E}_{f(x)}[\ln x]. \quad (B.2)$$

If x is gamma distributed as

$$f(x) = \text{Gamma}(x;\mu,\alpha), \quad (B.3)$$

the optimal $\widetilde{\ln x}$ writes

$$\widetilde{\ln x}^{*} = \psi(\mu)-\ln\alpha. \quad (B.4)$$

As shown in Fig. B1, $\ln x$ and $\psi(x)$ are very close to each other, especially when x becomes large, say $x\geq 5$. To simplify the expression and facilitate the calculation, we used $\ln x$ to approximate $\psi(x)$ in (B.4). Then the optimal $\widetilde{\ln x}$ is approximated as

$$\widetilde{\ln x}^{*} \approx \ln\mu-\ln\alpha = \ln\overline{x}. \quad (B.5)$$

As we took the first-order Taylor expansion for each variable iteratively, the overall lower-bound for the MLIB is tightened.

Secondly, we study the usage of Jensen's inequality in (24). As $\psi(x+y)$ is a concave function of x, for *any* x_0, we have

$$\mathbb{E}_{f(x)}[\psi(x+y)] \leq \mathbb{E}_{f(x)}[\psi(x_0+y)+\psi'(x_0+y)(x-x_0)]. \quad (B.6)$$

Similarly, the optimal x_0 that minimizes the RHS of (B.6) is

$$x_0^{*} = \mathbb{E}_{f(x)}[x] = \overline{x} = \frac{\mu}{\alpha}. \quad (B.7)$$

When we take $x_0^{*}=\overline{x}$, (B.6) is exactly the same as applying Jensen's inequality to $\psi(x+y)$. The same argument holds when we apply Jensen's inequality to $\ln x$.

Thus, Jensen's inequalities applied for $\psi(x+y)$ and $\ln x$ are optimal for tightening the lower-bound.

In summary, taking the first-order Taylor expansions around $\ln\overline{x}$ and applying Jensen's inequality are the optimal choices for tightening the lower-bound.

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, San Diego, CA, USA, 1990.
[2] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4–37.
[3] C.M. Bishop, Pattern Recognition and Machine Learning, 1st ed., Springer, New York, 2006.
[4] J.M. Bernardo, A. Smith, Bayesian Theory, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester, 2000.
[5] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (3) (1993) 803–821.
[6] Y. Ji, C. Wu, P. Liu, J. Wang, K.R. Coombes, Application of beta-mixture models in bioinformatics, Bioinform. Appl. Note 21 (2005) 2118–2122.
[7] Z. Ma, Non-Gaussian statistical models and their applications (Ph.D. thesis), KTH - Royal Institute of Technology, 2011.
[8] Z. Ma, A. Leijon, Bayesian estimation of beta mixture models with variational inference, IEEE Trans. Pattern Anal. Mach. Intell.. 33 (11) (2011) 2160–2173.
[9] Z. Ma, A.E. Teschendorff, A variational Bayes beta mixture model for feature selection in DNA methylation studies, J. Bioinform. Comput. Biol. 11 (4) (2013) 1350005 (19 pp.).
[10] N. Bouguila, D. Ziou, J. Vaillancourt, Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application, IEEE Trans. Image Process. 13 (11) (2004) 1533–1543.
[11] D. Blei, Probabilistic models of text and images (Ph.D. thesis), Univ. of California, Berkeley, 2004.
[12] W. Fan, N. Bouguila, D. Ziou, Variational learning for finite Dirichlet mixture models and applications, IEEE Trans. Neural Netw. Learn. Syst. 23 (5) (2012) 762–774.
[13] Z. Ma, A. Leijon, W.B. Kleijn, Vector quantization of LSF parameters with a mixture of Dirichlet distributions, IEEE Trans. Audio Speech Lang. Process. 21 (2013) 1777–1790.
[14] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, USA, 2000.
[15] D.M. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures, Bayesian Anal. 1 (2005) 121–144.
[16] P. Orbanz, Y.W. Teh, Bayesian nonparametric models, Encyclopedia of Machine Learning, 2010, pp 81-89.
[17] Z. Ghahramani, Bayesian non-parametrics and the probabilistic approach to modelling. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 371 (1984) (2013) 20110553.
[18] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul, An introduction to variational methods for graphical models, Mach. Learn. 37 (2) (1999) 183–233.
[19] T.S. Jaakkola, Tutorial on variational approximation methods, in: Advanced Mean Field Methods: Theory and Practice, MIT Press, Cambridge, MA, USA, 2000, pp. 129–159.

[20] T.S. Jaakkola, Tutorial on variational approximation methods, in: Advances in Mean Field Methods, MIT Press, Cambridge, MA, USA, 2001, pp. 129–159.
[21] D.M. Blei, J.D. Lafferty, Correlated topic models, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 2006, pp. 147—154.
[22] D.M. Blei, J.D. Lafferty, A correlated topic model of Science, Ann. Appl. Stat. 1 (2007) 17–35.
[23] M. Braun, J. McAuliffe, Variational inference for large-scale models of discrete choice, J. Am. Stat. Assoc. 105 (2010) 324–335.
[24] M. Hoffman, D. Blei, P. Cook, Bayesian nonparametric matrix factorization for recorded music, in: Proceedings of International Conference on Machine Learning, 2010.
[25] Z. Ma, A. Leijon, Modeling speech line spectral frequencies with Dirichlet mixture models, in: Proceedings of INTERSPEECH, 2010, pp. 2370–2373.
[26] M. Tanimoto, M.P. Tehrani, T. Fujii, T. Yendo, Free-viewpoint TV, IEEE Signal Process. Mag. 28 (1) (2011) 67–76.
[27] R. Ksantini, D. Ziou, B. Colin, F. Dubeau, Weighted pseudometric discriminatory power improvement using a Bayesian logistic regression model based on a variational method, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 253–266.
[28] P.J. Bickel, K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Pearson Prentice Hall, New Jersey, USA, 2007.
[29] J.A. Palmer, Relative convexity, Technical Report, UCSD, 2003.
[30] Z. Ma, A. Leijon, Beta mixture models and the application to image classification, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2009, pp. 2045 –2048.
[31] R.J. Connor, J.E. Mosimann, Concepts of independence for proportions with a generalization of the Dirichlet distribution, J. Am. Stat. Assoc. 64 (325) (1969) 194–206.
[32] P.K. Rana, Z. Ma, J. Taghia, M. Flierl, Multiview depth map enhancement by variational Bayes inference estimation of Dirichlet mixture models, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 1528–1532.
[33] K.K. Paliwal, W.B. Kleijn, Quantization of LPC parameters, in: Speech Coding and Synthesis, Elsevier, Amsterdam, The Netherlands, pp. 433–466, 1995.
[34] J. Lindblom, J. Samuelsson, Bounded support Gaussian mixture modeling of speech spectra, IEEE Trans. Speech Audio Process. 11 (1) (2003) 88–99.
[35] A.D. Subramaniam, B.D. Rao, PDF optimized parametric vector quantization of speech line spectral frequencies, IEEE Trans. Speech Audio Process. 11 (2) (2003) 130–142.
[36] S. Chatterjee, T.V. Sreenivas, Low complexity wideband LSF quantization using GMM of uncorrelated Gaussian mixtures, in: 16th European Signal Processing Conference (EUSIPCO 2008), 2008.
[37] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1, NASA STI/Recon Technical Report N 93, 1993, p. 27403.
[38] S. So, K.K. Paliwal, Empirical lower bound on the bitrate for the transparent memoryless coding of wideband LPC parameters, IEEE Signal Process. Lett. 13 (9) (2006) 569–572.
[39] A.A. Neath, J.E. Cavanaugh, The Bayesian information criterion: background, derivation, and applications, Wiley Interdiscip. Rev. Comput. Stat. 4 (2) (2012) 199–203.
[40] MPEG, Call for proposals on 3D video coding technology, Technical Report N12036, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, March 2011.
[41] H. Urey, K. Chellappan, E. Erden, P.A. Surman, State of the art in stereoscopic and autostereoscopic displays, Proc. IEEE 99 (4) (2011) 540–555.
[42] K. Müller, P. Merkle, T. Wiegand, 3-D video representation using depth maps, Proc. IEEE 99 (4) (2011) 643–656.
[43] M. Flierl, B. Girod, Multiview video compression, IEEE Signal Process. Mag. 24 (6) (2007) 66–76.
[44] A. Smolic, K. Müller, N. Stefanoski, J. Ostermann, A. Gotchev, G. Akar, G. Triantafyllidis, A. Koz, Coding algorithms for 3DTV–A survey, IEEE Trans. Circuits Syst. Video Technol. 17 (11) (2007) 1606–1621.
[45] C. Fehn, Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV, vol. 5291, SPIE, 2004, pp. 93–104.
[46] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vision 47 (2002) 7–42.
[47] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1124–1137.
[48] P. Felzenszwalb, D. Huttenlocher, Efficient belief propagation for early vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, Washington, DC, USA, 2004, pp. 261–268.
[49] P.K. Rana, M. Flierl, Depth consistency testing for improved view interpolation, in: Proceedings of the IEEE International Workshop on Multimedia Signal Processing, St. Malo, France, 2010, pp. 384–389.
[50] P.K. Rana, J. Taghia, M. Flierl, A variational Bayesian inference framework for multiview depth image enhancement, in: Proceedings of the IEEE International Symposium on Multimedia, Irvine, California, USA, 2012, pp. 183–190.
[51] G. Wyszecki, W. Stiles, Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd ed., John Wiley & Sons, New York, NY, USA, 2000.
[52] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, Y. Mori, Reference Softwares for Depth Estimation and View Synthesis, Technical Report M15377, ISO/IEC JTC1/SC29/WG11, Archamps, France, April 2008.
[53] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed., Cambridge University Press, Cambridge, UK, 2004.
[54] G. Wolberg, Digital Image Warping, 1st ed., IEEE Computer Society Press, Los Alamitos, CA, USA, 1994.
[55] L. McMillan, An image-based approach on three-dimensional computer graphics (Ph.D. thesis), University of North Carolina, Chapel Hill, 1997.
[56] W.R. Mark, Post-rendering 3D image warping: visibility, reconstruction, and performance for depth-image (Ph.D. thesis), University of North Carolina, 1999.
[57] M. Bertalmio, A. Bertozzi, G. Sapiro, Navier–Stokes, fluid dynamics, and image and video inpainting, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Kauai, HI, USA, 2001, pp. 355–362.
[58] MPEG, View Synthesis Software Manual, ISO/IEC JTC1/SC29/WG11, release 3.5, September 2009.

Zhanyu Ma has been an assistant Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2013. He received his M.Eng. degree in Signal and Information Processing from BUPT (Beijing University of Posts and Telecommunications), China, and his Ph.D. degree in Electrical Engineering from KTH (Royal Institute of Technology), Sweden, in 2007 and 2011, respectively. 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.

Pravin Kumar Rana received the M.Sc. degree in Physics with specialization in Electronics and Communication from Ranchi University, Ranchi, India, in 2004, and the M. Tech. degree in Earth System Science and Technology from Indian Institute of Technology (IIT) Kharagpur, India, in 2008. In 2008, he joined the School of Electrical Engineering at KTH Royal Institute of Technology, Stockholm, Sweden, where he is currently working towards the Ph.D. degree. His research area includes multiview video processing, 3D and free-viewpoint TV, and computer vision. He is a student member of the IEEE.

Jalil Taghia is currently pursuing a Ph.D. degree at the Communication Theory laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. His research interest includes statistical signal processing and machine learning including Bayesian inference, variational approximations, latent variable models in particular with audio applications.

Markus Flierl received the Ph.D. degree in electrical engineering from Friedrich Alexander University, Erlangen, Germany, in 2003. He is an Associate Professor at the Autonomic Complex Communication Networks, Signals and Systems (ACCESS) Linnaeus Center, School of Electrical Engineering, KTH - Royal Institute of Technology, Stockholm, Sweden. From 2005 to 2008, he was a Visiting Assistant Professor at the Max Planck Center for Visual Computing and Communication at Stanford University, Stanford, CA. He is the author of the book Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond and of more than 50 other scientific publications.

He is the recipient of the 2007 Visual Communications and Image Processing Young Investigator Award.

Arne Leijon is a Professor in Hearing Technology at the KTH (Royal Institute of Technology) Sound and Image Processing Lab, Stockholm, Sweden, since 1994. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these aids, based on psychoacoustic modelling of sensory information transmission and subjective sound quality. He received the M.S. degree in Engineering Physics in 1971, and a Ph.D. degree in Information Theory in 1989, both from Chalmers University of Technology, Gothenburg, Sweden.