J Sign Process Syst
DOI 10.1007/s11265-013-0769-8

Predictive Distribution of the Dirichlet Mixture Model by Local Variational Inference

Zhanyu Ma · Arne Leijon · Zheng-Hua Tan · Sheng Gao

Received: 4 October 2012 / Revised: 8 March 2013 / Accepted: 25 April 2013

Abstract In Bayesian analysis of a statistical model, the predictive distribution is obtained by marginalizing over the parameters with their posterior distributions. Compared to the frequently used point estimate plug-in method, the predictive distribution leads to a more reliable result in calculating the predictive likelihood of the new upcoming data, especially when the amount of training data is small. The Bayesian estimation of a Dirichlet mixture model (DMM) is, in general, not analytically tractable. In our previous work, we have proposed a global variational inference-based method for approximately calculating the posterior distributions of the parameters in the DMM analytically. In this paper, we extend our previous study for the DMM and propose an algorithm to calculate the predictive distribution of the DMM with the local variational inference (LVI) method. The true predictive distribution of the DMM is analytically intractable. By considering the concave property of the multivariate inverse beta function, we introduce an upper-bound to the true predictive distribution. As the global minimum of this upper-bound exists, the problem is reduced to seek an approximation to the true predictive distribution. The approximated predictive distribution obtained by minimizing the upper-bound is analytically tractable, facilitating the computation of the predictive likelihood. With synthesized data and real data evaluations, the good performance of the proposed LVI based method is demonstrated by comparing with some conventionally used methods.

Keywords Predictive distribution · Dirichlet mixture model · Bayesian inference · Local variational inference

Z. Ma (✉) · S. Gao
Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: mazhanyu@ieee.org

S. Gao
e-mail: gaosheng@bupt.edu.cn

A. Leijon
School of Electrical Engineering, KTH - Royal Institute of Technology, Stockholm, Sweden
e-mail: leijon@kth.se

Z.-H. Tan
Department of Electronic Systems, Aalborg University, Aalborg, Denmark
e-mail: zt@es.aau.dk

1 Introduction

Predicting the likelihood of new upcoming data is a fundamental problem in statistical modeling [1]. The ultimate goal of statistical modeling is to find a suitable distribution to describe the underlying distribution of the training data and apply this distribution properly to new data [1, 2]. One frequently used method is to estimate the parameters of the distribution by the maximum a posteriori (MAP) estimation [2–4] and then plug in the estimated parameters to get the distribution for the new upcoming data. However, when the amount of training data is small, the variances of the estimated parameters are large and the point estimate

Published online: 30 August 2013

may lead to high uncertainty and unreliability. An alternative solution is to take the uncertainty of the parameters into account and employ the predictive distribution. The predictive density of a new vector $\mathbf{x}$ given the training data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ is

$$f(\mathbf{x}|\mathbf{X}) = \int f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\theta}$ denotes the parameters in the likelihood function $f(\mathbf{x}|\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}|\mathbf{X})$ is the estimated posterior distribution of the parameters $\boldsymbol{\theta}$ given $\mathbf{X}$ (with hyperparameter omitted). The predictive distribution averages all the possible predictive probabilities for all the possible parameter values, which is actually the expected value of $f(\mathbf{x}|\boldsymbol{\theta})$ with respect to the posterior distribution. Hence, it is more reliable than a simple point estimate plug-in method. This predictive distribution, which is based on integrating out the posterior distribution of the parameters, is referred to as the "posterior predictive distribution" [5, 6]. Replacing the posterior distribution by the prior distribution, the so-called "prior predictive distribution" [5] can be obtained correspondingly. Usually, the prior distribution is assumed to be non-informative when we have no prior knowledge. As we focus on calculating the predictive likelihood of new data given the training data, we only consider the "posterior predictive distribution" and use the term "predictive distribution" to denote the "posterior predictive distribution" throughout this paper. It is worthy to note that, when the amount of training observations becomes sufficiently large, the posterior distribution tends to be concentrated over a small region of the parameter space [2]. In this case, the point estimate to the parameter (*e.g.*, the MAP estimate) could be considered to have a high degree of certainty[1] and the point estimate plug-in method can be used as a reasonable approximation to the predictive distribution as

$$f(\mathbf{x}|\mathbf{X}) \approx f\left(\mathbf{x}|\hat{\boldsymbol{\theta}}\right), \tag{2}$$

where $\hat{\boldsymbol{\theta}}$ is the point estimate. Figure 1 illustrates the concentration of the posterior distributions with different amounts of training data. Please note that this point estimate plug-in method is not accurate when we have a small amount of training data.

In statistical modeling, Gaussian distribution is the ubiquitous probability distribution used in statistics, since it has an analytically tractable probability density function (PDF) and analysis based on it can be derived in an explicit form [7, 8]. Furthermore, by the technique of mixture modeling [9–11], the corresponding Gaussian Mixture Model (GMM) can be used to approximate arbitrary probability distributions. However, in real life, not all the data we would like to model are Gaussian distributed [12, 13]. The Gaussian distribution has unbounded support, while some data are semi-bounded or bounded. For example, the digitalized image pixel values are bounded, the signal-to-noise ratio of wireless channel is always nonnegative (semi-bounded), and the line spectral frequency parameters are bounded and ordered. Therefore, they are non-Gaussian distributed. Applying GMM to describe the underlying distribution of these non-Gaussian distributed data would lead to a high model complexity, *e.g.*, many mixture components would be spent on describing the edge of the data space. In order to explicitly exploit the bounded/semi-bounded properties, some non-Gaussian distributions can be applied to efficiently model the underlying distribution of the non-Gaussian distributed data. Moreover, the consequent applications can also benefit from choosing the non-Gaussian distributions. Many studies demonstrated that the usage of non-Gaussian distributions is advantageous in applications where the data is not Gaussian distributed (see *e.g.*,[14–16]).

The Dirichlet distribution, among other non-Gaussian distributions, has been intensively studied and used to model data distribution in various applications, such as image processing [17, 18], multiview depth image enhancement [19, Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference, unpublished], speech coding [16, 20], and data mining [18, 21]. In addition to modeling the data's distribution directly, the Dirichlet distribution is also widely used to model the underlying distribution of the mixture weights in the mixture modeling framework [9, 10]. 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 [22–25]. In this paper, we only study the finite Dirichlet mixture model (DMM) and the work conducted can also be extended to the infinite mixture modeling case.

To model distribution with multimodality, the mixture modeling technique [9, 10] can be applied to get a DMM. To fit the DMM to the training data, the maximum likelihood (ML) estimation, which can be carried out by the expectation maximization (EM) algorithm [11], was proposed in [17, 20]. However, there are some general drawbacks in the ML estimation: 1) the estimated model might be overtraining to the data [11, 13]; 2) due to the integral expression in the multivariate-inverse-beta (MIB) function in the Dirichlet PDF, the maximization step involves numerical calculation [17, 20, 26], which is computationally costly; and 3) the EM based algorithm cannot decide the model

[1] In an extreme case, if the posterior distribution has no variance, the point estimate has absolute certainty.

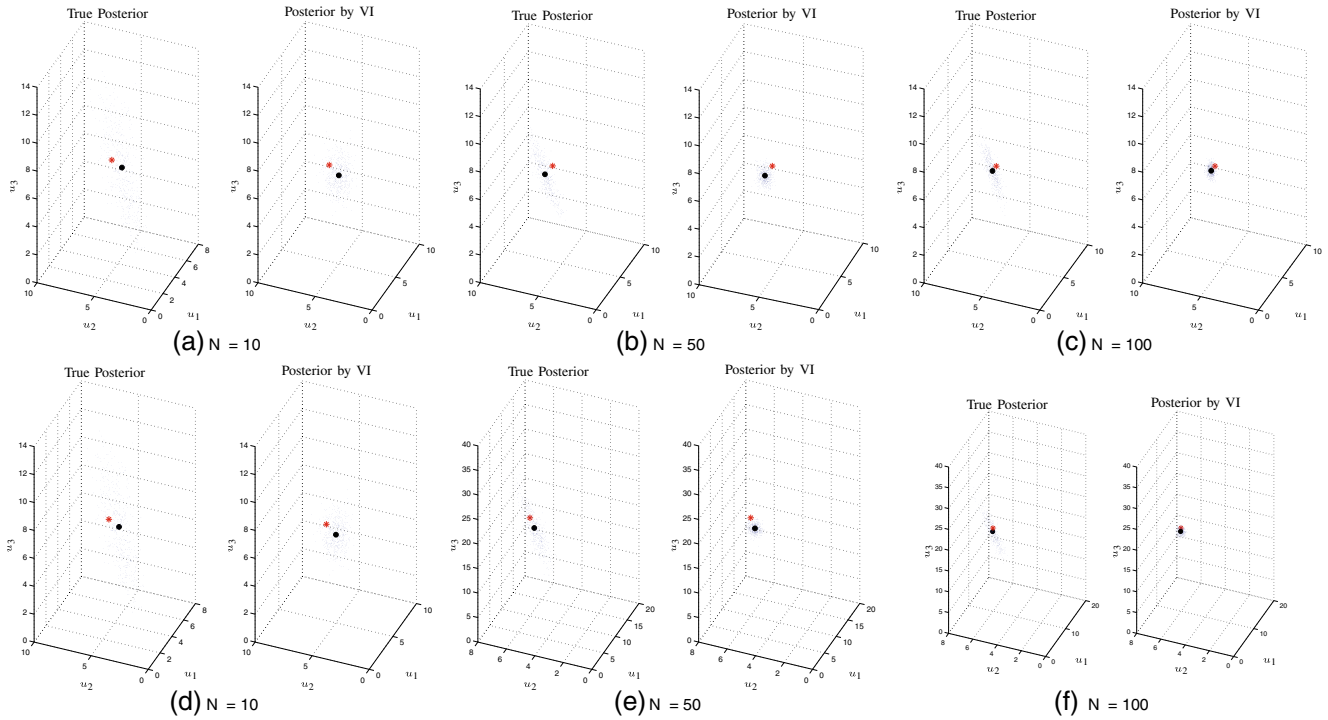

Figure 1 Comparisons of the true posterior distribution and the approximating one obtained by VI. The true posterior distribution was obtained by the rejection sampling method [2], where the reference sampling distribution is the Laplace approximation to the true one. Different amounts of training data 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 $\mathbf{u} = [10\ 6\ 20]^{\mathrm{T}}$. The mismatch between the true and the approximating posterior distributions is 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 training data increases.

complexity automatically by itself. To overcome these problems, Bayesian estimation of the generalized Dirichlet mixture model was proposed in [27]. With minimum message length criterion, the over-training problem could be avoided and the model complexity can be decided based on the data. However, the update strategy employed numerical calculation and, therefore, it is computationally costly, especially when dealing with high-dimensional data.

In order to obtain an analytically tractable solution, we proposed a Bayesian estimation method for the DMM in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished).[2] The proposed method utilized the variational inference (VI) framework. With the relative convexity[3] [29, 30] of the MIB function, we approximated the MIB function by its first-order Taylor expansion with respect to the logarithm of the variables. By the principle of the extended factorized approximation (EFA) [14, 31–33], an analytically tractable solution to the Bayesian estimation of the DMM was obtained and the parameter estimation was facilitated Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished). This VI-based method introduced some systematic bias, because the Dirichlet parameters were assumed to be mutually independent. However, with a sufficiently large amount of training data, the variance is small. As suggested in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished), the posterior mean can be used as the point estimate to the parameters. Another Bayesian estimation method for a single Dirichlet distribution, which was based on the expectation propagation (EP) framework [34, 35], was also

[2]There was another Bayesian estimation method proposed in [28]. However, the method introduced in [28] used the multiple lower-bounds (MLB) approximation to derive an analytically tractable solution. Different from [28], the method presented in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished) used the single lower-bound (SLB) approximation. As discussed in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished), the MLB approximation based solution cannot guarantee the convergency, while the SLB approximation based solution is more concise and can guarantee the convergency.

[3]If a function $f(x)$ is not convex in x but convex in $\ln x$, it is called "convex relative to" $\ln x$.

proposed in [36]. Different from the VI-based method, the EP-based method approximated the posterior distribution of the parameters in a Dirichlet distribution by a multivariate Gaussian distribution, which captured the correlations of the parameters but violated the nonnegativity of the parameters. The EP-based method proposed in [36] performed better than the VI-based method with a smaller amount of training data. The price for improvement is the numerical calculation employed in the moment-matching step. Moreover, the EP-based method was only proposed for a single Dirichlet distribution, not a mixture of Dirichlet distributions.

As mentioned above, the predictive distribution is more reliable than the point estimate plug-in method when predicting the likelihood of upcoming data. Thus, it is of interest to study the predictive distribution of the DMM. In the context of calculating the predictive distribution of the DMM, there are two challenging tasks: one is to obtain the posterior distributions of the parameters and the other is to derive an analytically tractable solution to approximately calculate the predictive distribution. The first problem has been addressed in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished), where the posterior distributions of the parameters in a Dirichlet distribution are approximated by a product of mutually independent gamma distributions. In this paper, we will focus on solving the second problem with the local variational inference (LVI) framework [2]. Unlike the VI framework, which can be referred to as the global VI (GVI) method, the LVI method seeks a bound on a subset of the variables in a model [2]. With the LVI framework, we have already proposed an approximation to the predictive distribution of a single beta distribution in [37]. The Dirichlet distribution is an extension of the beta distribution. The MIB function in the Dirichlet distribution is approximated by its first-order Taylor expansion. This expansion is an upper-bound of the MIB function, as the MIB function is jointly concave [38] with respect to all its variables. With this approximation and by the principle of the LVI method, the true predictive distribution is approximated by an analytically tractable expression, which is an upper-bound to the true one. It can be shown that the global minimum of this upper-bound exists. By minimizing this upper-bound and with normalization, we get an approximation to the true predictive distribution. The obtained approximation can be expressed in an analytically tractable form. With synthesized data and real data evaluations, the good performance of the proposed method is demonstrated, especially when the amount of training data is small.

The remaining parts of this paper are organized as follows: Section 2 describes the DMM. In Section 3, the Bayesian estimation of the DMM is introduced, from which we can obtain the point estimates to the parameters. With the obtained posterior distribution in Section 3, we approximately calculate the predictive distribution of the DMM in Section 4. The evaluations with synthesized data and real data are presented in Section 5. Finally, the conclusions are drawn in Section 6.

2 Dirichlet Mixture Model

Given a K dimensional vector $\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$, which contains only nonnegative elements and the summation of these elements equals one, the underlying distribution of $\mathbf{x}$ could be described by a Dirichlet distribution as[4]

$$\mathbf{Dir}\left(\mathbf{x};\mathbf{u}\right) = \frac{\Gamma\left(\sum_{k=1}^{K} u_k\right)}{\prod_{k=1}^{K}\Gamma\left(u_k\right)} \prod_{k=1}^{K} x_k^{u_k-1},\ u_k > 0, \tag{3}$$

where $\sum_{k=1}^{K} x_k = 1$, $\mathbf{u} = [u_1, \ldots, u_K]^{\mathrm{T}}$ 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 $u_k > 1$, $k = 1, \ldots, K$, it is unimodally distributed. This is a typical case in practical applications. Thus in this paper, we only study the Dirichlet distribution with all its parameters greater than one.

To model the multimodality of the data, the mixture modeling technique [9] is usually applied to build a DMM. With I mixture components, the PDF of a DMM can be represented, given a set of N *i.i.d.* observations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, as

$$f\left(\mathbf{X};\Pi,\mathbf{U}\right) = \prod_{n=1}^{N}\sum_{i=1}^{I}\pi_i \mathbf{Dir}\left(\mathbf{x}_n;\mathbf{u}_i\right), \tag{4}$$

where $\Pi = [\pi_1, \ldots, \pi_I]^{\mathrm{T}}$ and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_I]$ are the parameter sets. As the gamma function in the multivariate-inverse-beta (MIB) function $\frac{\Gamma\left(\sum_{k=1}^{K} u_k\right)}{\prod_{k=1}^{K}\Gamma\left(u_k\right)}$ is expressed by an integral form, the maximization step in the EM algorithm [17, 20] cannot be carried out by an analytically tractable solution. Some greedy search algorithms, *e.g.*, the Newton–Raphson algorithm, are required to calculate the stationary points numerically. Although the computational load in the Newton-Raphson algorithm has been significantly reduced by some matrix tricks [26], the EM algorithm is still computationally costly.

[4] 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.

3 Bayesian Estimation of the Dirichlet Mixture Model by the Global Variational Inference

As belonged to the exponential family, the conjugate prior of the Dirichlet distribution writes

$$f\left(\mathbf{u};\boldsymbol{\beta}_0,\nu_0\right)=\frac{1}{C\left(\boldsymbol{\beta}_0,\nu_0\right)}\left[\frac{\Gamma\left(\sum_{k=1}^{K}u_k\right)}{\prod_{k=1}^{K}\Gamma(u_k)}\right]^{\nu_0}e^{-\boldsymbol{\beta}_0^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_K)}, \tag{5}$$

where $\boldsymbol{\beta}_0=[\beta_{1_0},\ldots,\beta_{K_0}]^{\mathrm{T}}$ and ν_0 are the hyperparameters in the prior distribution. $C\left(\boldsymbol{\beta}_0,\nu_0\right)$ is the normalization factor. $\mathbf{1}_m$ denotes an m dimensional vector with all elements equal to one. Combining (3) and (5) together and by Bayes' rule, the posterior distribution of the parameters $\mathbf{u}$ can be obtained as

$$\begin{aligned}&f\left(\mathbf{u}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N\right)\\&=\frac{\mathbf{Dir}\left(\mathbf{X}|\mathbf{u}\right)f\left(\mathbf{u};\boldsymbol{\beta}_0,\nu_0\right)}{\int\mathbf{Dir}\left(\mathbf{X}|\mathbf{u}\right)f\left(\mathbf{u};\boldsymbol{\beta}_0,\nu_0\right)d\mathbf{u}}\\&=\frac{1}{C\left(\boldsymbol{\beta}_N,\nu_N\right)}\left[\frac{\Gamma\left(\sum_{k=1}^{K}u_k\right)}{\prod_{k=1}^{K}\Gamma(u_k)}\right]^{\nu_N}e^{-\boldsymbol{\beta}_N^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_K)},\end{aligned} \tag{6}$$

where $\boldsymbol{\beta}_N=\boldsymbol{\beta}_0-\ln\mathbf{X}\times\mathbf{1}_N$ and $\nu_N=\nu_0+N$ are the hyperparameters in the posterior distribution. Obviously, some sufficient statistics, *e.g.*, the mean, the covariance matrix, cannot be calculated by an analytically tractable form, so that it is not convenient to use the posterior distribution derived in Eq. 6.

In order to seek easy-to-use conjugate prior and posterior distributions, a Bayesian estimation method, which is based on the global variational inference (GVI) framework, was proposed in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished). By assuming the elements in $\mathbf{u}$ are mutually independent, the prior distribution in Eq. 5 was approximately factorized into a product of several gamma distributions as Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished)

$$\begin{aligned}f\left(\mathbf{u};\boldsymbol{\beta}_0,\nu_0\right)&\approx f\left(\mathbf{u};\boldsymbol{\mu}_0,\boldsymbol{\alpha}_0\right)\\&=\prod_{k=1}^{K}\mathbf{Gam}\left(u_k;\mu_{k_0},\alpha_{k_0}\right),\end{aligned} \tag{7}$$

where $\mu_{k_0},\alpha_{k_0},\ k=1,2,\ldots,K$ are the hyperparameters in the prior distribution and

$$\mathbf{Gam}\left(u;\mu,\alpha\right)=\frac{\alpha^{\mu}}{\Gamma(\mu)}u^{\mu-1}e^{-\alpha u}. \tag{8}$$

With the relative convexity [29], the logarithm of the MIB function was approximated by its first-order Taylor expansion, which is a lower-bound to it. Afterwards, with Jensen's inequality, the variational objective function (variational lower-bound in the GVI framework) was lower-bounded by an auxiliary function. Using the principles of the extended factorized approximation (EFA) method [14, 33], the posterior distribution of each variable, *i.e.*, $f(u_k|\mathbf{X})$, was shown to be gamma distributed as

$$f\left(u_k|\mathbf{X}\right)=\mathbf{Gam}\left(u_k|\mathbf{X};\mu_k^*,\alpha_k^*\right), \tag{9}$$

where μ_k^* and α_k^* are the optimal posterior hyperparameters obtained by an analytically tractable solution. Finally, the posterior distribution of $\mathbf{u}$ was approximated as

$$\begin{aligned}f\left(\mathbf{u}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N\right)&\approx f\left(\mathbf{u}|\mathbf{X};\boldsymbol{\mu}^*,\boldsymbol{\alpha}^*\right)\\&=\prod_{k=1}^{K}\mathbf{Gam}\left(u_k|\mathbf{X};\mu_k^*,\alpha_k^*\right).\end{aligned} \tag{10}$$

The assumption of mutually independence among the elements in $\mathbf{u}$ violates the correlation, this assumption introduced some systematic bias due to the EFA framework. However, when the amount of observations increases, both the true posterior distribution and the approximating one are concentrated in a small region of the parameter space, and then the effect of this bias is small. Figure 1 shows comparisons between the true and the approximating posterior distributions. Therefore, when the amount of training data is sufficiently large, the posterior means, *i.e.*, $\overline{u}_k=\mathbf{E}[u_k]=\mu_k^*/\alpha_k^*$, can be taken as point estimates to the parameters, as suggested by Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished).

4 Predictive Distribution of the Dirichlet Mixture Model

When calculating the predictive likelihood of new data with a small amount of training data, the predictive distribution in Eq. 1 is more reliable than the point estimate plug-in method in Eq. 2. The true predictive distribution of the Dirichlet distribution writes

$$f(\mathbf{x}|\mathbf{X})=\int\mathbf{Dir}(\mathbf{x}|\mathbf{u})f\left(\mathbf{u}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N\right)d\mathbf{u}. \tag{11}$$

Since the true posterior distribution is not feasible in practice, the approximating posterior distribution obtained in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished) can be used to approximately calculate the predictive distribution.

The predictive distribution of the Dirichlet distribution, with the approximated posterior distribution in Eq. 10, is

$$\begin{aligned} f(\mathbf{x}|\mathbf{X}) &= \int \mathbf{Dir}(\mathbf{x}|\mathbf{u}) f\left(\mathbf{u}|\mathbf{X};\boldsymbol{\mu}^*,\boldsymbol{\alpha}^*\right) d\mathbf{u} \\ &= \int \frac{\Gamma\left(\sum_{k=1}^K u_k\right)}{\prod_{k=1}^K \Gamma(u_k)} \prod_{k=1}^K x_k^{u_k-1} \\ &\quad \times \prod_{k=1}^K \frac{\left(\alpha_k^*\right)^{\mu_k^*}}{\Gamma\left(\mu_k^*\right)} u_k^{\mu_k^*-1} e^{-\alpha_k^* u_k} d\mathbf{u} \\ &= \int \cdots \int \frac{\Gamma\left(\sum_{k=1}^K u_k\right)}{\prod_{k=1}^K \Gamma(u_k)} \\ &\quad \times x_1^{u_1-1} \frac{\left(\alpha_1^*\right)^{\mu_1^*}}{\Gamma\left(\mu_1^*\right)} u_1^{\mu_1^*-1} e^{-\alpha_1^* u_1} \\ &\quad \cdots \\ &\quad \times x_K^{u_K-1} \frac{\left(\alpha_K^*\right)^{\mu_K^*}}{\Gamma\left(\mu_K^*\right)} u_K^{\mu_K^*-1} e^{-\alpha_K^* u_K} du_1 \cdots du_K. \end{aligned} \tag{12}$$

Apparently, the integration in Eq. 12 cannot be calculated by an analytically tractable form. One way to calculate it numerically is to employ the sampling method, which requires generation of a huge amount of samples from the posterior distribution and is computationally inefficient. In the following paragraph, we will focus on deriving an analytically tractable expression to approximately calculate the predictive distribution.

4.1 Local Variational Inference

The global variational inference (GVI) framework used in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished) can be considered as a "global" method because it uses bound to approximate the variational objective function in terms of all the variables. As an alternative method to the GVI, an "local" approach involves finding bounds on a subset of the variables or a part of the objective function [2]. This method is referred to as the local variational inference (LVI) framework. The purpose of using the LVI framework to introduce a bound is to simplify the resulting distribution. For multiple variables, this local approximation can be applied in turn until a tractable approximation is obtained [2]. Suppose we would like to evaluate an analytically intractable integration

$$\mathrm{F} = \int f(x) g(x) dx, \tag{13}$$

where $f(x)$ is the PDF of x and $g(x)$ is a function of x. If there exits an auxiliary function $h(x,\sigma)$ such that

1. $h(x,\sigma) \geq g(x)$
2. $\int f(x)h(x,\sigma)dx$ can be calculated explicitly,

then the integration in Eq. 13 can be approximated, by the principle of the LVI, as

$$\mathrm{F} \leq \mathrm{G}(\sigma) = \int h(x,\sigma) g(x) dx. \tag{14}$$

Thus, an upper-bound to F, i.e., $\mathrm{G}(\sigma)$, is obtained. As $\mathrm{G}(\sigma)$ is only a function of σ, the true value F can be approximated by finding the optimal σ^*, which minimizes $\mathrm{G}(\sigma)$ as

$$\sigma^* = \arg\min_{\sigma} \mathrm{G}(\sigma). \tag{15}$$

In general, the optimized value $\mathrm{G}(\sigma^*)$ is not exactly the same as the true one. However, the LVI based method facilitates the calculation by little loss of accuracy [37].

4.2 Approximation to a Single Dirichlet Distribution

In this section, we firstly study the concave property of the multivariate-inverse-beta (MIB) function and then derive an upper-bound to the true predictive distribution of a single Dirichlet distribution. Afterwards, we discuss the existence of the global minimum and approximate it by a simple but efficient form. Finally, the approximation to the true predictive density will be expressed by an analytically tractable expression.

4.2.1 Concavity of the Multivariate-Inverse-Beta Function

Theorem 1 *The logarithm of the MIB function is*

$$\mathbf{Q}(\mathbf{u}) = \ln \frac{\Gamma\left(\sum_{k=1}^K u_k\right)}{\prod_{k=1}^K \Gamma(u_k)} = \ln\Gamma\left(\sum_{k=1}^K u_k\right) - \sum_{k=1}^K \ln\Gamma(u_k). \tag{16}$$

This function is jointly concave with respect to all its variables $\mathbf{u} = [u_1, u_2, \ldots, u_K]^{\mathrm{T}}$.

Proof Concavity of $\mathrm{Q}(\mathbf{u})$ The elements in the Hessian matrix of $\mathbf{Q}(\mathbf{u})$ writes

$$H_{ij} = \begin{cases} \psi'\left(\sum_{k=1}^K u_k\right) - \psi'(u_i) & i = j \\ \psi'\left(\sum_{k=1}^K u_k\right) & i \neq j \end{cases}, \tag{17}$$

where $\psi'(\cdot)$ is the derivative of the digamma function

$$\psi'(x) = \frac{\partial \psi(x)}{\partial x} \text{ and } \psi(x) = \frac{\partial \ln\Gamma(x)}{\partial x}.$$

In matrix form, the Hessian matrix can be expressed as

$$\mathbf{H} = \mathbf{z}\mathbf{z}^{\mathrm{T}} - \mathbf{A}, \tag{18}$$

where

$$\mathbf{z} = [z_1, z_2, \ldots, z_K]^{\mathrm{T}},$$
$$z_1 = z_2 = \cdots = z_K = \sqrt{\psi'\left(\sum_{k=1}^{K} u_k\right)}, \tag{19}$$

and

$$\mathbf{A} = \begin{bmatrix} \psi'(u_1) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \psi'(u_K) \end{bmatrix}. \tag{20}$$

For any $k = 1, 2, \ldots, K$, the upper-left $k \times k$ sub-matrix of $\mathbf{H}$ writes

$$\mathbf{H}_k = \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}} - \mathbf{A}_k, \tag{21}$$

where $\mathbf{z}_k$ contains the first k elements in $\mathbf{z}$ and $\mathbf{A}_k$ is the upper-left $k \times k$ sub-matrix of $\mathbf{A}$.

The concavity of $\mathbf{Q}(\mathbf{u})$ can be proved as follows:

1. When $k = 1$, we have

$$\mathbf{H}_1 = \psi'\left(\sum_{k=1}^{K} u_k\right) - \psi'(u_1) < 0 \tag{22}$$

because $\psi'(x)$ is a monotonously nonnegative decreasing function of x and $u_k > 1$.

2. When $k = 2$, the determinant of $\mathbf{H}_2$ is

$$\begin{aligned} |\mathbf{H}_2| &= |-\mathbf{A}_2|\left[1 - \mathbf{z}_2^{\mathrm{T}}(\mathbf{A}_2)^{-1}\mathbf{z}_2\right] \\ &= \prod_{i=1}^{k} \psi'(u_i)\left\{1 - \psi'\left(\sum_{k=1}^{K} u_k\right) \times \left[\frac{1}{\psi'(u_1)} + \frac{1}{\psi'(u_2)}\right]\right\}. \end{aligned} \tag{23}$$

In Eq. 23, we used the matrix calculus trick [39]

$$\left|\mathbf{A} + \mathbf{x}\mathbf{y}^{\mathrm{T}}\right| = |\mathbf{A}|\left(1 + \mathbf{y}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{x}\right), \tag{24}$$

where $\mathbf{A}$ is a square $m \times m$ matrix and $\mathbf{x}$, $\mathbf{y}$ are vectors of size $m \times 1$.

Before deriving the sign of $|\mathbf{H}_2|$, let us study the sign of the function

$$\mathbf{G}(u_1, u_2) = \psi'(u_1)\psi'(u_2) - \left[\psi'(u_1) + \psi'(u_2)\right]\psi'(u_1 + u_2). \tag{25}$$

By fixing $u_1 = \tilde{u}_1$, it is observed that $\mathbf{G}(\tilde{u}_1, u_2)$ is monotonously decreasing function of u_2. Since $\lim_{u_2 \to \infty} \mathbf{G}(\tilde{u}_1, u_2) = 0$, we conclude that $\mathbf{G}(\tilde{u}_1, u_2) > 0$ for any u_2. For the same reasoning, by fixing $u_2 = \tilde{u}_2$, we also have $\mathbf{G}(u_1, \tilde{u}_2) > 0$ for any u_1. Combining these two arguments together, we have

$$\mathbf{G}(u_1, u_2) > 0 \tag{26}$$

for any u_1 and u_2.

From Eq. 26, the following inequality holds as

$$1 - \psi'(u_1 + u_2)\left[\frac{1}{\psi'(u_1)} + \frac{1}{\psi'(u_2)}\right] > 0. \tag{27}$$

Recall that $\psi'(x)$ is a monotonously decreasing nonnegative function, we have

$$1 - \psi'\left(\sum_{k=1}^{K} u_k\right)\left[\frac{1}{\psi'(u_1)} + \frac{1}{\psi'(u_2)}\right] > 0. \tag{28}$$

Substituting (28) into (23), it is easily to get

$$|\mathbf{H}_2| = \prod_{k=1}^{K} \underbrace{\psi'(u_k)}_{>0} \underbrace{\left\{1 - \psi'\left(\sum_{k=1}^{K} u_k\right)\left[\frac{1}{\psi'(u_1)} + \frac{1}{\psi'(u_2)}\right]\right\}}_{>0} > 0 \tag{29}$$

$$\begin{aligned} f(\mathbf{x}|\mathbf{X}) \leq & \int \cdots \int \frac{\Gamma\left(\sum_{k=1}^{K} \tilde{u}_k\right)}{\prod_{k=1}^{K} \Gamma(\tilde{u}_k)} \\ & \times e^{\sum_{k=1}^{K}\left[\psi\left(\sum_{k=1}^{K} \tilde{u}_k\right) - \psi(\tilde{u}_k)\right](u_k - \tilde{u}_k)} \\ & \times x_1^{u_1 - 1} \frac{(\alpha_1^*)^{\mu_1^*}}{\Gamma(\mu_1^*)} u_1^{\mu_1^* - 1} e^{-\alpha_1^* u_1} \\ & \times \cdots \times x_K^{u_K - 1} \frac{(\alpha_K^*)^{\mu_K^*}}{\Gamma(\mu_K^*)} u_K^{\mu_K^* - 1} e^{-\alpha_K^* u_K} du_1 \cdots du_K \\ = & \frac{\Gamma\left(\sum_{k=1}^{K} \tilde{u}_k\right)}{\prod_{k=1}^{K} \Gamma(\tilde{u}_k)} \times e^{-\sum_{k=1}^{K} \tilde{u}_k\left[\psi\left(\sum_{k=1}^{K} \tilde{u}_k\right) - \psi(\tilde{u}_k)\right]} \\ & \times \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k \Gamma(\mu_k^*)} \int e^{-u_k\left[\alpha_k^* - \ln x_k - \psi\left(\sum_{k=1}^{K} \tilde{u}_k\right) + \psi(\tilde{u}_k)\right]} u_k^{\mu_k^* - 1} du_k \\ \triangleq & f_{\mathrm{upp}}(\mathbf{x}|\mathbf{X}). \end{aligned} \tag{30}$$

3. When $k \geq 3$ and k is odd, the determinant of $\mathbf{H}_k$ is

$$\begin{aligned} |\mathbf{H}_k| &= |-\mathbf{A}_k|\left[1 - \mathbf{z}_k^{\mathrm{T}}(\mathbf{A}_k)^{-1}\mathbf{z}_k\right] \\ &= -\prod_{i=1}^{k} \psi'(u_i)\left\{1 - \psi'\left(\sum_{k=1}^{K} u_k\right)\left[\sum_{i=1}^{k} \frac{1}{\psi'(u_i)}\right]\right\}. \end{aligned} \tag{31}$$

The inequality in Eq. 27 could also be written as

$$\frac{1}{\psi'(u_1)}+\frac{1}{\psi'(u_2)}<\frac{1}{\psi'(u_1+u_2)}. \tag{32}$$

Using this inequality iteratively, we get

$$\begin{aligned}\sum_{i=1}^{k}\frac{1}{\psi'(u_i)} &= \frac{1}{\psi'(u_1)}+\frac{1}{\psi'(u_2)}+\frac{1}{\psi'(u_3)}\\ &\quad+\cdots+\frac{1}{\psi'(u_k)}<\frac{1}{\psi'(u_1+u_2)}\\ &\quad+\frac{1}{\psi'(u_3)}+\cdots+\frac{1}{\psi'(u_k)}\\ &<\frac{1}{\psi'(u_1+u_2+u_3)}+\cdots+\frac{1}{\psi'(u_k)}\\ &\cdots\\ &<\frac{1}{\psi'\left(\sum_{i=1}^{k}u_i\right)}\\ &<\frac{1}{\psi'\left(\sum_{k=1}^{K}u_k\right)},\end{aligned}$$

which can also be written as

$$\psi'\left(\sum_{k=1}^{K}u_k\right)\left[\sum_{i=1}^{k}\frac{1}{\psi'(u_i)}\right]<1. \tag{33}$$

Substituting (33) into (31), we conclude that

$$|\mathbf{H}_k|<0,\ k\geq 3 \text{ and } k \text{ is odd}. \tag{34}$$

4. Similarly, when $k\geq 3$ and k is even, we have

$$|\mathbf{H}_k|>0,\ k\geq 3 \text{ and } k \text{ is even}. \tag{35}$$

As all the kth order leading principal minors of $\mathbf{H}$ are negative when k is odd and positive when k is even, it is sufficient to prove that $\mathbf{H}$ is negative-definite. Therefore, $\mathbf{Q}(\mathbf{u})$ is jointly concave with respect to $\mathbf{u}$ [38]. □

The logarithm of the MIB function has been proved to be concave, which indicates the following relation:

$$\begin{aligned}\ln\frac{\Gamma\left(\sum_{k=1}^{K}u_k\right)}{\prod_{k=1}^{K}\Gamma(u_k)} &\leq \ln\frac{\Gamma\left(\sum_{k=1}^{K}\tilde{u}_k\right)}{\prod_{k=1}^{K}\Gamma(\tilde{u}_k)}\\ &\quad+\sum_{k=1}^{K}\left[\psi\left(\sum_{k=1}^{K}\tilde{u}_k\right)-\psi(\tilde{u}_k)\right](u_k-\tilde{u}_k),\end{aligned} \tag{36}$$

where $\tilde{u}_k,\ k=1,2,\ldots,K$ is any point from the posterior distribution. Taking the exponential of both sides, we have

$$\begin{aligned}\frac{\Gamma\left(\sum_{k=1}^{K}u_k\right)}{\prod_{k=1}^{K}\Gamma(u_k)} &\leq \frac{\Gamma\left(\sum_{k=1}^{K}\tilde{u}_k\right)}{\prod_{k=1}^{K}\Gamma(\tilde{u}_k)}\\ &\quad\times e^{\sum_{k=1}^{K}\left[\psi\left(\sum_{k=1}^{K}\tilde{u}_k\right)-\psi(\tilde{u}_k)\right](u_k-\tilde{u}_k)},\end{aligned} \tag{37}$$

which means the RHS of Eq. 37 is an upper-bound to the MIB function.

4.2.2 Upper-bound of the Predictive Distribution

To approximate the predictive distribution in Eq. 12, the LVI method [2] can be applied to introduce an upper-bound. Substituting (37) into (12) and with some mathematics, an upper-bound to the predictive distribution is obtained as in Eq. 30. Denoting

$$\mathbf{G}\left(x_k,\tilde{\mathbf{u}}\right)=\alpha_k^*-\ln x_k-\psi\left(\sum_{k=1}^{K}\tilde{u}_k\right)+\psi\left(\tilde{u}_k\right), \tag{38}$$

where $k=1,2,\ldots,K$, the integrand in each integration of Eq. 30 is recognized to have the same form of the gamma distribution (in terms of u_k). By this, the integrations in Eq. 30 is then calculated as

$$\int e^{-u_k\mathbf{G}(x_k,\tilde{\mathbf{u}})}u_k^{\mu_k^*-1}du_k=\begin{cases}\frac{\Gamma(\mu_k^*)}{[\mathbf{G}(x_k,\tilde{\mathbf{u}})]^{\mu_k^*}} & \mathbf{G}\left(x_k,\tilde{\mathbf{u}}\right)>0\\ \infty & \mathbf{G}\left(x_k,\tilde{\mathbf{u}}\right)\leq 0\end{cases}. \tag{39}$$

With the assumption that $\mathbf{G}\left(x_k,\tilde{\mathbf{u}}\right)>0$ for any k, the finite upper-bound of the predictive distribution is written in an analytically tractable form as

$$\begin{aligned}f_{\text{upp}}(\mathbf{x}|\mathbf{X}) &= \frac{\Gamma\left(\sum_{k=1}^{K}\tilde{u}_k\right)}{\prod_{k=1}^{K}\Gamma\left(\tilde{u}_k\right)}\times e^{-\sum_{k=1}^{K}\tilde{u}_k\left[\psi\left(\sum_{k=1}^{K}\tilde{u}_k\right)-\psi(\tilde{u}_k)\right]}\\ &\quad\times\prod_{k=1}^{K}\frac{\left(\alpha_k^*\right)^{\mu_k^*}}{x_k\left[\mathbf{G}\left(x_k,\tilde{\mathbf{u}}\right)\right]^{\mu_k^*}}.\end{aligned} \tag{40}$$

For any data vector $\mathbf{x}$, the upper-bound to the predictive distribution is only a function of $\tilde{\mathbf{u}}$, because $\alpha_k^*,\ \mu_k^*,\ k=1,2,\ldots,K$ are the hyperparameters in the posterior distribution and fixed once the Bayesian estimation of the parameters is done.

4.2.3 Existence of the Upper-bound's Global Minimum

The derived upper-bound in Eq. 42 is equal to the predictive distribution in Eq. 12 only if the Taylor expansion is tight to the original function. This requires the posterior distribution of u_k to be a Dirac delta function and has zero variance. When increasing the amount of training data, the posterior distribution gets concentrated in a small region of the parameter space. The upper-bound is asymptotically tight to the true function when the amount of observations goes to infinity. Thus, there exists a systematic gap between (42) and (12) and we need to find the minimum

value of $f_{\text{upp}}(\mathbf{x}|\mathbf{X})$ to reduce the gap. In order to do this, the following constrained optimization problem must be solved:

$$\begin{aligned} &\min_{\tilde{\mathbf{u}}} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) \\ &\text{s.t. } \mathbf{G}(x_k, \tilde{\mathbf{u}}) > 0 \end{aligned}, \tag{41}$$

where $f_{\text{upp}}(\mathbf{x}|\mathbf{X})$ is rewritten as

$$f_{\text{upp}}(\mathbf{x}|\mathbf{X}) = \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k} \tag{42}$$

and

$$\begin{aligned} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) = & \frac{\Gamma\left(\sum_{k=1}^{K} \tilde{u}_k\right)}{\prod_{k=1}^{K} \Gamma(\tilde{u}_k)} \times e^{-\sum_{k=1}^{K} \tilde{u}_k \left[\psi\left(\sum_{k=1}^{K} \tilde{u}_k\right) - \psi(\tilde{u}_k)\right]} \\ & \times \prod_{k=1}^{K} \frac{1}{\mathbf{G}(x_k, \tilde{\mathbf{u}})^{\mu_k^*}}. \end{aligned} \tag{43}$$

The objective function $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ is not strictly convex or concave, but we will now show that a global minimum of $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ exists, for any given $\mathbf{x}$.

Firstly, we study the properties of the constraints. By considering $\tilde{u}_j$ is the only changing variable in $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ and fixing the rest $\tilde{u}_i$, $i = 1, 2, \ldots, K$, $i \neq j$, it can be shown that $\mathbf{G}(x_j, \tilde{\mathbf{u}})$ is a monotonously increasing function of $\tilde{u}_j$ while $\mathbf{G}(x_i, \tilde{\mathbf{u}})$ is monotonously decreasing in terms of $\tilde{u}_j$. This statement comes from the fact that $\psi(x)$ is a monotonously increasing function of x.

Then, the constraints can be categorized into two groups. The increasing constraint is

$$\begin{aligned} \mathbf{G}\left(x_j, \tilde{\mathbf{u}}\right) &= \mathbf{G}\left(x_j, \tilde{\mathbf{u}}_{\setminus j}, \tilde{u}_j\right) \ ^{5} \\ &= \alpha_j^* - \ln x_j - \psi\left(\sum_{k=1, k\neq j}^{K} \tilde{u}_k + \tilde{u}_j\right) + \psi\left(\tilde{u}_j\right) \\ &> 0. \end{aligned} \tag{44}$$

There exists an point u_j^l so that $\mathbf{G}\left(x_j, \tilde{\mathbf{u}}_{\setminus j}, u_j^l\right) = 0$. Then we have $\tilde{u}_j > u_j^l$ to satisfy this constraint. Meanwhile, the decreasing constraint, for $i = 1, 2, \ldots, K$, $i \neq j$, is

$$\begin{aligned} \mathbf{G}(x_i, \tilde{\mathbf{u}}) &= \mathbf{G}\left(x_i, \tilde{\mathbf{u}}_{\setminus j}, \tilde{u}_j\right) \\ &= \alpha_i^* - \ln x_i - \psi\left(\sum_{k=1, k\neq j}^{K} \tilde{u}_k + \tilde{u}_j\right) + \psi\left(\tilde{u}_i\right) \\ &> 0. \end{aligned} \tag{45}$$

When $\mathbf{G}\left(x_i, \tilde{\mathbf{u}}_{\setminus j}, u_j^{r_i}\right) = 0$, $i = 1, 2, \ldots, K$, $i \neq j$, we have $\tilde{u}_j < \min_i u_i^{r_i}$ to satisfy these $K - 1$ decreasing constraints. Thus, for the known $\tilde{\mathbf{u}}_{\setminus j}$, there exists an open interval for $\tilde{u}_j$ as $u_j^l < \tilde{u}_j < \min_i u_j^{r_i}$. Hence, for any element $\tilde{u}_k$ in $\tilde{\mathbf{u}}$, when $\tilde{\mathbf{u}}_{\setminus k}$ is fixed, there exists an open interval to satisfy all the K constraints. The above discussions show that, to satisfy the constraint $\mathbf{G}(x_k, \tilde{\mathbf{u}}) > 0$ for any k, $\tilde{\mathbf{u}}$ should fall in an K dimensional space defined as

$$\mathbb{F}^K = \left\{\tilde{\mathbf{u}} : \bigcap_{k=1}^{K} \mathbf{G}(x_k, \tilde{\mathbf{u}}) > 0\right\}. \tag{46}$$

It is noteworthy that $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) = +\infty$, if $\tilde{\mathbf{u}}$ does not fall in $\mathbb{F}^K$ (see Eq. 39).

Secondly, we consider the case when $\tilde{\mathbf{u}}$ falls on the boundary of $\mathbb{F}^K$. The boundary of $\mathbb{F}^K$ is defined by $\left\{\tilde{\mathbf{u}} : \bigcup_{k=1}^{K} \mathbf{G}(x_k, \tilde{\mathbf{u}}) = 0\right\}$. From Eq. 39, it can be shown that $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) = +\infty$ if $\tilde{\mathbf{u}}$ is on the boundary of $\mathbb{F}^K$. Thus, when $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ is finite, $\tilde{\mathbf{u}}$ is not on the boundary. Then the constraint $\mathbf{G}(x_k, \tilde{\mathbf{u}}) > 0$ for any k can be extended to $\mathbf{G}(x_k, \tilde{\mathbf{u}}) \geq 0$, without any influence on the minimization operation in Eq. 41.

As the maximum of $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ is $+\infty$ and $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ is a continuous nonnegative function, the minimum value of $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ exists. Assuming that $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ has M stationary points, each of which satisfies $\nabla \mathbf{P}\left(\mathbf{x}, \tilde{\mathbf{u}}_m^*\right) = 0$, $m = 1, 2, \ldots, M$. Then $\min_{\tilde{\mathbf{u}}} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ can be found by comparing all the $\mathbf{P}\left(\mathbf{x}, \tilde{\mathbf{u}}_m^*\right)$ and choosing the smallest as

$$\min_{\tilde{\mathbf{u}}} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) = \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}^*). \tag{47}$$

Therefore, the optimal solution is

$$\tilde{\mathbf{u}}^* = \arg\min_{\tilde{\mathbf{u}}_m^*} \mathbf{P}\left(\mathbf{x}, \tilde{\mathbf{u}}_m^*\right). \tag{48}$$

4.2.4 Approximation to $\min_{\tilde{\mathbf{u}}} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$

Some numerical algorithms can be applied to obtain $\tilde{\mathbf{u}}^*$. However, no matter what kind of method it is, the result of the optimization procedure depends on $\mathbf{x}$. Thus, for any new $\mathbf{x}$, the optimization problem has a different optimal solution and must be solved again. This is computationally costly and practically infeasible in Bayesian learning applications.

An alternative way to solve this problem is to make the optimal solution in Eq. 41 independent of $\mathbf{x}$, *i.e.*, to approximate the optimal point $\tilde{\mathbf{u}}^*$ by a value that does not depend on $\mathbf{x}$. Then, given the estimated posterior distribution, the global minimum of $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ is fixed for any $\mathbf{x}$.

To this end, we use the posterior mean of $\mathbf{u}$ from the posterior distribution in Eq. 10 to approximate the optimal solution in Eq. 41. There are three reasons for this choice:

1. $f_{\text{upp}}(\mathbf{x}|\mathbf{X})$ was obtained by using the first-order Taylor expansion as an upper-bound. It is observed that [14; 33, Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference, unpublished], when taking the Taylor expansion around the expected value of the argument, this bound is tight.
2. In principle, the value of $\tilde{\mathbf{u}}$ can be arbitrarily selected in the domain of $\mathbf{u}$. However, as $\mathbf{u}$ is distributed according

[5] $\tilde{\mathbf{u}}_{\setminus j}$ denotes all the elements in $\tilde{\mathbf{u}}$ except $\tilde{u}_j$.

to the posterior distribution in Eq. 10, it is reasonable to use a representative point from the posterior distribution to approximate the optimal solution. As each element in $\mathbf{u}$ is gamma distributed and the gamma distribution is unimodal, the posterior mean can be considered as a representative point. Moreover, when the posterior distribution concentrates in a small region, this approximation becomes more accurate.

3. The posterior mean does not depend on $\mathbf{x}$, which facilitates the calculation.

Thus, it is reasonable to take the posterior mean to approximate the true optimal solution as

$$\min_{\tilde{\mathbf{u}}} \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}) \approx \mathbf{P}(\mathbf{x}, \bar{\mathbf{u}}). \tag{49}$$

Indeed, this approximation approach will lead to some unknown bias. Substituting this approximation into (42) and with normalization (see next section for details), the approximation to the predictive distribution is obtained. Table 1 shows the Kullback–Leibler (KL) divergences[6] from the true predictive distribution to the approximating one. It can be observed that both methods (one is based on greedy search (47) and the other is based on the posterior mean (49)) can approximate the true predictive distribution properly. The predictive distribution approximate by using the posterior mean ($f_{\text{appx}}^{\text{post.}}(\mathbf{x}|\mathbf{X})$ in Eq. 52) is slightly worse than the one obtained by greedy search ($f_{\text{appx}}^{\text{search}}(\mathbf{x}|\mathbf{X})$ in Eq. 53). This is because of using the posterior mean to approximate the optimal solution. However, as argued above, the variance introduced by using the posterior mean approximation decreases when the amount of observations increases. Table 2 shows the comparison of the KL divergence from the predictive distribution approximated by greedy search to the one approximated by using the posterior mean. The differences between these two methods are very small. As using the posterior mean can facilitate the calculation significantly with little loss of accuracy, it is acceptable in practice.

Table 1 Comparisons of the KL divergences ($\times 10^4$) of the true predictive distribution (in Eq. 11) from the approximating one. $f_{\text{appx}}^{\text{search}}(\mathbf{x}|\mathbf{X})$ is the predictive distribution approximated by using (47) and $f_{\text{appx}}^{\text{post.}}(\mathbf{x}|\mathbf{X})$ is the one approximated by using (49).

Method	$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
$f_{\text{appx}}^{\text{search}}(\mathbf{x}\|\mathbf{X})$	2.035	0.764	0.550	0.404	0.358
$f_{\text{appx}}^{\text{post.}}(\mathbf{x}\|\mathbf{X})$	9.453	1.406	0.575	0.429	0.368
	$N = 60$	$N = 70$	$N = 80$	$N = 90$	$N = 100$
$f_{\text{appx}}^{\text{search}}(\mathbf{x}\|\mathbf{X})$	0.296	0.304	0.295	0.284	0.289
$f_{\text{appx}}^{\text{post.}}(\mathbf{x}\|\mathbf{X})$	0.304	0.318	0.314	0.299	0.294

The true predictive distribution was obtained by the importance sampling method [2]. Different amounts of data were generated from a Dirichlet distribution with parameter $\mathbf{u} = [10\ 6\ 20]^{\mathrm{T}}$. The means of 20 rounds of simulations are reported. Similar performance can be obtained by some other parameter settings and we show only one example here

4.2.5 Final Approximation

In the above section, we discussed the global minimum of $\mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}})$ and proposed an approximation to the global minimum by using the posterior mean of $\mathbf{u}$. Substituting (49) into (42), the global minimum of the upper-bound can be expressed as

$$\begin{aligned} \min_{\tilde{\mathbf{u}}} f_{\text{upp}}(\mathbf{x}|\mathbf{X}) &= \min_{\tilde{\mathbf{u}}} \mathbf{P}\left(\mathbf{x}, \tilde{\mathbf{u}}\right) \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k} \\ &\approx \mathbf{P}\left(\mathbf{x}, \widehat{\mathbf{u}}\right) \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k}. \end{aligned} \tag{50}$$

Table 2 Comparisons of the KL divergences ($\times 10^4$) from $f_{\text{appx}}^{\text{search}}(\mathbf{x}|\mathbf{X})$ to $f_{\text{appx}}^{\text{post.}}(\mathbf{x}|\mathbf{X})$.

$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
8.240	0.389	0.367	0.130	0.066
$N = 60$	$N = 70$	$N = 80$	$N = 90$	$N = 100$
0.032	0.025	0.023	0.011	0.001

The data were generated from a Dirichlet distribution with parameter $\mathbf{u} = [10\ 6\ 20]^{\mathrm{T}}$. The reported values are the means of 20 rounds of simulations. Similar performance can be obtained by some other parameter settings and we show only one example here

Algorithm 1 Predictive Distribution of the DMM

1. Run the Bayesian estimation method proposed in Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference (unpublished), obtain the mixture weights Π and the hyperparameters $\boldsymbol{\alpha}^*$ $\boldsymbol{\mu}^*$ from the posterior distribution.
2. For each mixture component, calculate the posterior means from the hyperparameters as $\bar{\mathbf{u}}_i = \boldsymbol{\mu}_i^* \oslash \boldsymbol{\alpha}_i^*$.[7]
3. Calculate the normalization factor $\mathbf{C}_i$ for each mixture component.
4. For any upcoming $\mathbf{x}$, the predictive distribution can be approximately calculated by Eq. 55.

[6] The KL divergence from $f(x)$ to $g(x)$ is calculated as $\mathrm{KL}(f\|g) = \int f(x) \ln \frac{f(x)}{g(x)} dx$

[7] $\oslash$ is the element-wise division.

Since $\min_{\tilde{\mathbf{u}}} f_{\text{upp}}(\mathbf{x}|\mathbf{X})$ is unnormalized, we can calculated the normalization factor

$$\begin{aligned}\mathbf{C}_{\text{post.}} &= \int \min_{\tilde{\mathbf{u}}} f_{\text{upp}}(\mathbf{x}|\mathbf{X})d\mathbf{x} \\ &= \int \mathbf{P}(\mathbf{x}, \hat{\mathbf{u}})d\mathbf{x} \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k}. \end{aligned} \tag{51}$$

Thus, the approximation to the true predictive distribution is finally obtained as

$$\begin{aligned} f(\mathbf{x}|\mathbf{X}) &\approx f_{\text{appx}}^{\text{post.}}(\mathbf{x}|\mathbf{X}) \\ &= \frac{1}{\mathbf{C}_{\text{post.}}} \times \mathbf{P}(\mathbf{x}, \hat{\mathbf{u}}) \times \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k}. \end{aligned} \tag{52}$$

If we use the true global minimum in Eq. 47, the approximation writes

$$\begin{aligned} f(\mathbf{x}|\mathbf{X}) &\approx f_{\text{appx}}^{\text{search}}(\mathbf{x}|\mathbf{X}) \\ &= \frac{1}{\mathbf{C}_{\text{search}}} \times \mathbf{P}(\mathbf{x}, \tilde{\mathbf{u}}^*) \times \prod_{k=1}^{K} \frac{(\alpha_k^*)^{\mu_k^*}}{x_k}. \end{aligned} \tag{53}$$

It is noteworthy that the *only* numerical calculation required is to calculated the normalization factor $\mathbf{C}$. Once this normalization factor is calculated, the predictive likelihood of *any* $\mathbf{x}$ can be obtained in an analytically tractable form. Compared to the true predictive distribution in Eq. 11, which requires numerical calculation for each new $\mathbf{x}$, this method indeed reduces computational cost.

4.3 Approximation to the Predictive Distribution of the Dirichlet Mixture Model

The posterior distribution of the DMM can be obtained by the Bayesian estimation method proposed in Ma et al., Bayesian estimation of Dirichlet mixture model with variational inference (unpublished). The predictive likelihood of an upcoming data given an estimated DMM is

$$f(\mathbf{x}|\mathbf{X}) = \sum_{i=1}^{I} \pi_i \int \mathbf{Dir}(\mathbf{x}|\mathbf{u}_i) f(\mathbf{u}_i|\mathbf{X})d\mathbf{u}_i. \tag{54}$$

With the LVI framework and using the approximation derived in Eq. 52, the predictive distribution of the DMM can be approximated as

$$\begin{aligned} f(\mathbf{x}|\mathbf{X}) &\approx f_{\text{appx}}^{\text{LVI}}(\mathbf{x}|\mathbf{X}) \\ &= \sum_{i=1}^{I} \pi_i \left[\frac{1}{\mathbf{C}_i} \times \mathbf{P}(\mathbf{x}, \widehat{\mathbf{u}}_i) \times \prod_{k=1}^{K} \frac{(\alpha_{ki}^*)^{\mu_{ki}^*}}{x_k} \right]. \end{aligned} \tag{55}$$

The algorithm of the LVI based predictive distribution for the DMM is presented in Algorithm 1.

5 Experimental Results and Discussion

The proposed LVI based method is evaluated with both synthesized and real data. In the synthesized data evaluation, different amounts of data are generated from a known model and the LVI based method is compared with two conventional point estimate plug-in methods, namely the GVI based method and the ML based method. In the real data evaluation, the proposed method is applied to classify Electroencephalogram (EEG) signal.

5.1 Synthesized Data Evaluation

In statistical modeling, there are several ways to approximate the predictive distribution. The proposed LVI based method can approximately calculate the predictive distribution. In addition to this method, there are two other conventionally used approximations, which are all based on the point estimate plug-in method. One approximation is based on the ML estimation to the parameters [20], which can be written as

$$f(\mathbf{x}|\mathbf{X}) \approx f_{\text{appx}}^{\text{ML}}(\mathbf{x}|\mathbf{X}) = \sum_{i=1}^{I} \pi_i \mathbf{Dir}\left(\mathbf{x}; \widehat{\mathbf{u}}_i^{\text{ML}}\right), \tag{56}$$

where $\widehat{\mathbf{u}}_i^{\text{ML}}$ is the ML estimate to the ith mixture component. The other approximation, which is based on the GVI framework (Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference, unpublished) and uses the posterior mean as the point estimates, writes

$$f(\mathbf{x}|\mathbf{X}) \approx f_{\text{appx}}^{\text{GVI}}(\mathbf{x}|\mathbf{X}) = \sum_{i=1}^{I} \pi_i \mathbf{Dir}\left(\mathbf{x}; \overline{\mathbf{u}}_i^{\text{GVI}}\right), \tag{57}$$

where $\overline{\mathbf{u}}_i^{\text{GVI}}$ is the posterior mean of $\mathbf{u}$ in the ith mixture component.

The comparisons between the true and the approximating predictive distributions are illustrated in Fig. 2. The predictive likelihood difference (ΔPL) is the absolute differences between the true predictive distribution and the approximating one as

$$\Delta\text{PL} = \left| f(\mathbf{x}|\mathbf{X}) - f_{\text{appx}}^{k}(\mathbf{x}|\mathbf{X}) \right|, \ k \in \{\text{LVI, GVI, ML}\}.$$

It can be observed that the LVI based method can approximate the true predictive distribution more accurate than both the GVI based and the ML based methods, especially when the amount of training data is small (*e.g.*, $N = 10$). For different amounts of observations, the ΔPL via the LVI based method is smaller than both of the other two referred methods. Moreover, as the amount of training data increases, all the three methods show improving approximation performance.

To compare the LVI based method with the GVI based and ML based methods qualitatively, we also evaluated

the KL divergences from the true predictive distribution to the approximating one. Different amounts of data were generated from a known Dirichlet distribution. The KL divergences were calculated numerically by replacing the integration operation with the summation operation. Table 3 shows the comparisons. As expected, the LVI based method yields significantly smaller KL divergences than the other two methods, especially when the amount of training data is small. This is because the predictive distribution obtained by integration marginalizes over the uncertainty of the parameter and, therefore, leads to a robust prediction.

In the synthesized data evaluation, we used a three dimensional Dirichlet distribution purely for the purpose of an easy visualization. Similar performance can be obtained by other parameter settings. Moreover, since the true posterior distribution of the DMM cannot be obtained analytically (only the posterior distribution to a single Dirichlet distribution can be obtained by Eq. 6), the comparisons were made based on a single Dirichlet distribution and the LVI based method performs superior to both the GVI based and the ML based methods. As shown in Eq. 54, the predictive distribution of the DMM is the weighted sum of several single predictive distributions. Thus, when approximating the predictive distribution of DMM, the LVI based method potentially permits better performance than the other two referred methods.

5.2 Real Data Evaluation

Classification of the Electroencephalogram (EEG) signal is a challenging task in the design of brain-computer interface (BCI) systems [40]. The EEG signal, which can be acquired non-invasively, is a recording of the electrical activity along the scalp while a person is imagining a kind of action. For

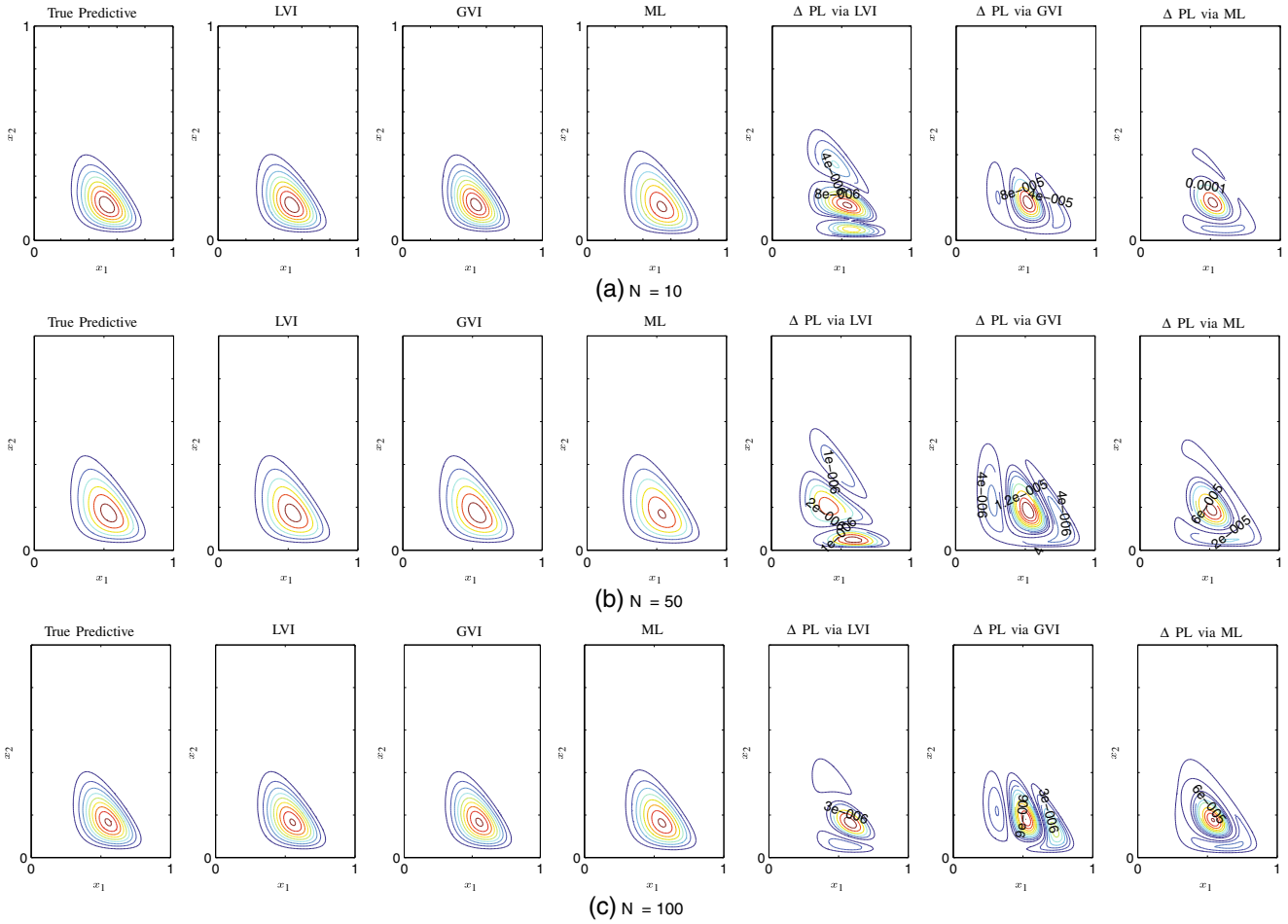

Figure 2 Comparisons of different approximations to the true predictive distribution. The true predictive distribution was obtained numerically by the importance sampling method [2]. Different amounts of training data were generated from the Dirichlet distribution $\mathbf{Dir}(\mathbf{x}; \mathbf{u})$ with known $\mathbf{u} = [3\ 5\ 8]^{\mathrm{T}}$. As the degrees of freedom equals 2, we have $x_3 = 1 - x_1 - x_2$ and omitted axis x_3 in the figure. For the convenience of visualization, these figures show the top-down view of the three dimensional predictive distributions and their differences. The mismatch is illustrated by the absolute difference (ΔPL) between the true and the approximating predictive density. The difference becomes smaller as the amount of training data increases.

Table 3 KL divergences ($\times 10^4$) from the true predictive distribution to the approximating one.

Method	$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
LVI	2.12	0.62	0.30	0.23	0.17
GVI	48.94	12.12	5.60	3.09	1.87
ML	512.85	268.06	138.41	83.85	98.55
	$N = 60$	$N = 70$	$N = 80$	$N = 90$	$N = 100$
LVI	0.13	0.12	0.11	0.09	0.08
GVI	1.30	0.91	0.69	0.54	0.39
ML	73.33	69.42	60.77	59.46	55.72

They were evaluated with the data generated from $\mathbf{Dir}(\mathbf{x}; \mathbf{u})$, $\mathbf{u} = [3\ 5\ 8]^{\mathrm{T}}$

different actions, EEG signals differ so that the type of actions can be estimated from the EEG signals. Thus, to classify the EEG signals properly is an essential part of a BCI system [41, 42].

The most frequently used feature in the EEG classification is the marginalized discrete wavelet transform (mDWT) coefficients [43, 44]. For each channel (*i.e.*, recording position on the scalp), the EEG signals were recorded separately and the mDWT coefficients were extracted. The mDWT coefficients from a single channel has the following properties: 1) all the coefficients are nonnegative and 2) the summation of the coefficients equals one [42]. To utilize these properties explicitly, a classifier based on the super-Dirichlet mixture model (sDMM) [45] was proposed in [42], where the classifier's evaluation was based on the well-known BCI competition III [46]. During the EEG signal recording, a subject performed two imagined movements: 1) the left small finger movement and 2) the tongue movement. In the end, two classes of EEG signals were obtained. The brain activities were recorded from 8×8 ECoG platinum electrode grid which was placed on the contralateral (right) motor cortex. In total, we have 64 channels of signals. 278 trials were recorded as the labeled training set and 100 trials were recorded as the unlabeled test set. In both the training set and test set, the data are evenly recorded for each imaginary movement. It is unclear which channels are more relevant for the imaginary task than the others [47] and the signals recorded from irrelevant channels may be noisy for the classification task [41]. As suggested in [42], the Fisher ratio (FR), which presents how strongly a channel correlates with class labels $\{-1, +1\}$, is applied to select the relevant channels. The FR is calculated as [48]

$$\mathrm{FR}(m) = \max_{\mathbf{d}} \frac{\mathbf{d}^{\mathrm{T}} \left[\boldsymbol{\mu}(m)_{+1} - \boldsymbol{\mu}(m)_{-1}\right] \left[\boldsymbol{\mu}(m)_{+1} - \boldsymbol{\mu}(m)_{-1}\right]^{\mathrm{T}} \mathbf{d}}{\mathbf{d}^{\mathrm{T}} \left[\Sigma(m)_{+1} + \Sigma(m)_{-1}\right] \mathbf{d}}, \tag{58}$$

Table 4 Comparisons of the average classification rates (in %) of the EEG signal over the top 25 channels.

Method	$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
LVI	52.15	52.97	52.68	52.81	53.00
GVI	52.01	52.12	51.89	52.17	52.42
	$N = 60$	$N = 70$	$N = 80$	$N = 90$	$N = 100$
LVI	53.61	53.89	53.71	53.80	53.85
GVI	53.58	53.76	53.62	53.78	53.73
	$N = 110$	$N = 120$	$N = 130$	$N = 139$ (All)	
LVI	53.78	53.72	53.69	53.88	
GVI	53.61	53.65	53.66	53.72	

N denotes the amount of training observations for each class which were randomly selected from the training set. The means of 20 simulation rounds are reported

where $\boldsymbol{\mu}(m)_j$ and $\Sigma(m)_j$, $m = 1, \ldots, 64$, $j \in \{+1, -1\}$ are the mean and the covariance matrix of class j in channel m, respectively. $\mathbf{d}$ is a vector with the same size of $\boldsymbol{\mu}(m)_j$. The channels with larger FRs are preferable. The reported results in [42] showed that the sDMM based classifier performed best when combining the top 21 or 24 channels that were selected by the FRs.

Thus, in this paper, we use only the top 25 channels, instead of all the 64 channels, to evaluate the proposed LVI based method. Moreover, as the ML estimation based predictive distribution performed far worse than either the LVI or the GVI based method (see Table 3), we only compare the LVI based and the GVI based methods. To highlight the advantage of the integration based predictive distribution, we train the classifier with different amounts of data from the training set. All the data from the test set are used for classification. Firstly, we studied the single channel based classification. The DMM was used to model the distribution of one class of EEG signal from a selected channel and estimated by the Bayesian estimation method (Ma et al., Bayesian estimation of Dirichlet mixture model with variational inference, unpublished). Afterwards, the proposed LVI based method and the GVI based method (Ma et al., Bayesian estimation of Dirichlet mixture model with variational inference, unpublished) were applied respectively to carry out the classification task. The average classification rates for all the 25 selected channels are listed in Table 4. It can be observed that the LVI based method yield better classification rate than the GVI based method for a wide range of amounts of observations. When the amount of observation increases, the classification performance is improved, although not monotonously. Secondly, we took a similar approach as [42] to carry out the classification task on more than one channel by the sDMM based classifier. For the

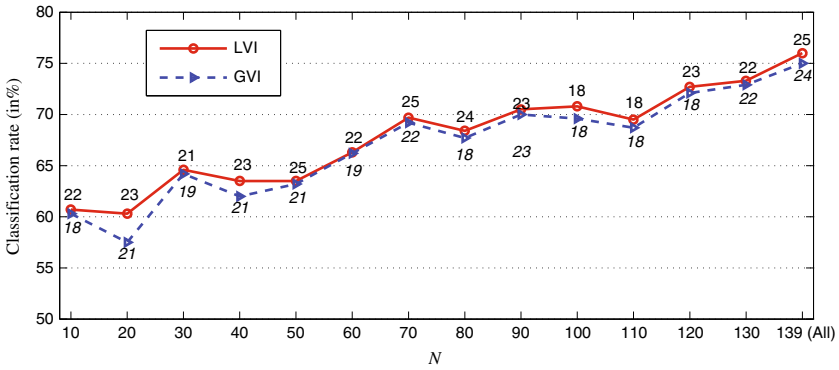

Figure 3 Classification rates of the EEG signal based on multiple channel combination for different amount of training data. The numbers above the *solid line* are the top L channels with which the LVI based method performs the best. The numbers below the *dash line* are the top L channels with which the GVI based method performs the best.

mDWT coefficients from L channels,[8] the PDF of a single super-Dirichlet distribution is

$$\begin{aligned}\mathbf{sDir}(\mathbf{x}_{\text{sup}};\mathbf{u}_{\text{sup}}) &= \prod_{l=1}^{L}\mathbf{Dir}(\mathbf{x}_l;\mathbf{u}_l)\\ &= \prod_{l=1}^{L}\frac{\Gamma\left(\sum_{k=1}^{K}u_{lk}\right)}{\prod_{k=1}^{K}\Gamma(u_{lk})}\prod_{k=1}^{K}x_{lk}^{u_{lk}-1},\end{aligned}\tag{59}$$

where

$$\mathbf{x}_{\text{sup}}=\begin{bmatrix}\mathbf{x}_1\\ \vdots\\ \mathbf{x}_L\end{bmatrix}\quad\text{and}\quad\mathbf{u}_{\text{sup}}=\begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_l\end{bmatrix}.\tag{60}$$

As a super-Dirichlet distribution is a cascade version of several single Dirichlet distribution [45], the Bayesian estimation method for DMM (Ma et al., Bayesian estimation of Dirichlet mixture model with varitional inference, unpublished) can be easily extended and applied to the sDMM. Similarly, the proposed LVI based predictive distribution in this paper can also be easily extended for sDMM. Figure 3 shows the classification rates obtained by multiple channels combination. The LVI based method performs better than the GVI based method for different amounts of training data. We believe this is because of the advantage of the integration-based predictive distribution.

6 Conclusion

The predictive distribution is, in general, more reliable than the point estimate plug-in method when calculating the predictive likelihood for new data, especially with a smaller amount of training observations. To approximately calculate the predictive distribution of the Dirichlet mixture model (DMM), we applied the local variational inference (LVI) framework to get an upper-bound to the multivariate inverse beta (MIB) function, by using the concavity of the MIB function. Then the predictive distribution is upper-bounded by an analytically tractable expression. The global minimum of this upper-bound expression was shown to exist and an efficient approximation was applied to calculate the global minimum. Finally, the predictive distribution of the DMM was approximated by an analytically tractable form, which facilitates the calculation of the predictive likelihood of the upcoming data. With synthesized and real data evaluations, the proposed LVI based method was superior in performance to the conventionally used global variational inference method and the maximum likelihood method.

[8] Here, the dimensionalities of the mDWT coefficients are the same for all the channels.

References

1. Bjørnstad, J.F. (1990). Predictive likelihood: a review. *Statistical Science*, *5*, 242–254.
2. Bishop, C.M. (2006). *Pattern recognition and machine learning*. New York: Springer.
3. Sorenson, H.W. (1980). *Parameter estimation: principles and problems*. New York: Marcel Dekker.
4. Kamen, E.W., & Su, J. (1999). *Introduction to optimal estimation, ser. Advanced textbooks in control and signal processing*. London: Springer.
5. Gelman, A., Meng, X.-L., Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.
6. Sinharay, S., & Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, *111*, 209–221.
7. Patel, J.K., & Read, C.B. (1996). *Handbook of the normal distribution, ser. Statistics, textbooks and monographs*. Marcel Dekker.

8. Jain, A.K., Duin, R.P.W., Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 4–37.
9. McLachlan, G., & Peel, D. (2000). *Finite mixture models, ser. Wiley series in probability and statistics: applied probability and statistics*. Wiley.
10. Figueiredo, M.A.T., & Jain, A.K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 381–396.
11. McLachlan, G.J., & Krishnan, T. (2008). *The EM algorithm and extensions, ser. Wiley series in probability and statistics*. Wiley-Interscience.
12. Banfield, J.D., & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.
13. Ma, Z. (2011). *Non-Gaussian statistical models and their applications*. Ph.D. dissertation, US-AB, Stockholm: KTH - Royal Institute of Technology.
14. Ma, Z., & Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2160–2173.
15. Atapattu, S., Tellambura, C., Jiang, H. (2011). A mixture Gamma distribution to model the SNR of wireless channels. *IEEE Transactions on Wireless Communications*, *10*(12), 4193–4203.
16. Ma, Z., Leijon, A., Kleijn, W.B. (2013). Vector quantization of LSF parameters with a mixture of Dirichlet distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(9), 1777–1790.
17. Bouguila, N., Ziou, D., Vaillancourt, J. (2004). Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application. *IEEE Transactions on Image Processing*, *13*(11), 1533–1543.
18. Blei, D.M. (2004). *Probabilistic models of text and images*. Ph.D, dissertation. University of California, Berkeley.
19. Rana, P.K., Ma, Z., Taghia, J., Flierl, M. (2013). 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)*.
20. Ma, Z., & Leijon, A. (2010). Modeling speech line spectral frequencies with Dirichlet mixture models. In *Proceedings of INTERSPEECH* (pp. 2370–2373).
21. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
22. Blei, D.M., & Jordan, M.I. (2005). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*, 121–144.
23. Orbanz, P., & Teh, Y.W. (2010). Bayesian nonparametric models. *Encyclopedia of Machine Learning*, 88–89.
24. Orbanz, P. (2010). Construction of nonparametric Bayesian models from parametric Bayes equations. In *Advances in neural information processing systems*.
25. Ghahramani, Z. (2012). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, *371*.
26. Minka, T.P. (2003). Estimating a Dirichlet distribution. *Annals of Physics*, *2000*(8), 1–13.
27. Bouguila, N., & Ziou, D. (2007). High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(10), 1716–1731.
28. Fan, W., Bouguila, N., Ziou, D. (2012). Variational learning for finite Dirichlet mixture models and applications. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(5), 762–774.
29. Palmer, J.A. (2003). Relative convexity. ECE Dept., UCSD Tech. Rep.
30. Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*, 17–35.
31. Jaakkola, T.S., & Jordan, M.I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*, 25–37.
32. Jaakkola, T.S. (2001). Tutorial on variational approximation methods. In M. Opper & D. Saad (Eds.), *Advances in mean field methods* (pp. 129–159). Cambridge: MIT Press.
33. Hoffman, M., Blei, D., Cook, P. (2010). Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the international conference on machine learning*.
34. Minka, T.P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 362–369).
35. Minka, T.P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. dissertation. Massachusetts Institute of Technology.
36. Ma, Z. (2012). Bayesian estimation of the Dirichlet distribution with expectation propagation. In *Proceeding of the 20th European signal processing conference* (pp. 689–693).
37. Ma, Z., & Leijon, A. (2011). Approximating the predictive distribution of the beta distribution with the local variational method. In *Proceedings of IEEE international workshop on machine learning for signal processing* (pp. 1–6).
38. Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
39. Brookes, M. (2013). The matrix reference manual. Available online: http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html. Accessed 9 Aug 2013.
40. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, *4*(2), R1.
41. Prasad, S., Tan, Z.-H., Prasad, R., Cabrera, A.F., Gu, Y., Dremstrup, K. (2011). Feature selection strategy for classification of single-trial EEG elicited by motor imagery. In *International symposium on wireless personal multimedia communications (WPMC)* (pp. 1–4).
42. Ma, Z., Tan, Z.-H., Prasad, S. (2012). EEG signal classification with super-Dirichlet mixture model. In *Proceedings of IEEE statistical signal processing workshop* (pp. 440–443).
43. Subasi, A. (2007). EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, *32*(4), 1084–1093.
44. Farina, D., Nascimento, O.F., Lucas, M.F., Doncarli, C. (2007). Optimization of wavelets for classification of movement-related cortical potentials generated by variation of force-related parameters. *Journal of Neuroscience Methods*, *162*, 357–363.
45. Ma, Z., & Leijon, A. (2011). Super-Dirichlet mixture models using differential line spectral frequences for text-independent speaker identification. In *Proceedings of INTERSPEECH* (pp. 2349–2352).
46. BCI competition III. http://www.bbci.de/competition/iii.
47. Lal, T.N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., Scholkopf, B. (2004). Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1003–1010.
48. Malina, W. (1981). On an extended fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3*(5), 611–614.

Zhanyu Ma is 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 2012-2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include statistical modeling and machine learning related topics with a focus on applications in speech processing, image processing, biomedical signal processing, and bioinformatics.

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 concernsapplied 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.

Zheng-Hua Tan received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark, which he joined in May 2001. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise robust speech processing, multimedia signal and information processing, multimodal human-computer interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He is an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing. He has served/serves as a program co-chair, area and session chair, tutorial speaker and committee member in many major international conferences.

Sheng Gao is an Assistant Professor of Beijing University of Posts and Telecommunications (BUPT) since 2012. He received his bachelor and Master degree from BUPT in 2003 and 2006 respectively. He got his Ph.D. degree from Universite Pierre et Marie CURIE (Paris 6), France in 2011. His research interests include machine learning, data mining, information recommendation, social network analysis, etc. He has published over 20 academic papers on world-wide famous journals or conferences including ISI, WWW, CIKM, ECML, etc.