BAYESIAN ESTIMATION OF THE DIRICHLET DISTRIBUTION WITH EXPECTATION PROPAGATION

Zhanyu Ma

KTH - Royal Institute of Technology
School of Electrical Engineering
Sound and Image Processing Lab.
Stockholm, Sweden

zhanyu@kth.se

ABSTRACT

As a member of the exponential family, the Dirichlet distribution has its conjugate prior. However, since the posterior distribution is difficult to use in practical problems, Bayesian estimation of the Dirichlet distribution, in general, is not analytically tractable. To derive practically easily used prior and posterior distributions, some approximations are required to approximate both the prior and the posterior distributions so that the conjugate match between the prior and posterior distributions holds and the obtained posterior distribution is easy to be employed. To this end, we approximate the distribution of the parameters in the Dirichlet distribution by a multivariate Gaussian distribution, based on the expectation propagation (EP) framework. The EP-based method captures the correlations among the parameters and provides an easily used prior/posterior distribution. Compared to recently proposed Bayesian estimation based on the variation inference (VI) framework, the EP-based method performs better with a smaller amount of observed data and is more stable.

Index Terms— Dirichlet distribution, Bayesian estimation, Expectation Propagation, Variational Inference

1. INTRODUCTION

In parametric statistical modeling study, parameter estimation plays an important role [1]. With the observed data, we assume that the observations are following a specified distribution and estimate the parameters to describe the probability density function (PDF) of this distribution. There are several ways to fit the parameters. The method of finding suitable values of the parameters by maximizing the likelihood function is named the maximum likelihood estimation (MLE). If we treat the parameters as random variables, the Bayesian estimation method can be applied to obtain the distributions of the parameters. The Bayesian estimation has several advantages over the MLE: 1) the Bayesian estimation not only provides some representative values (*e.g.*, the mode, the mean) but also describes the distributions of the parameters and 2) the Bayesian estimation can prevent the overfitting problem, which is in general a drawback of the MLE. Generally speaking, Bayesian estimation is more reliable than the MLE, especially when the amount of observed data is small.

The Gaussian distribution is the frequently used probability distribution in statistics. However, not all the data we would like to model are Gaussian distributed [2], due to the natural properties of the data. For example, the digitalized image pixel values are bounded within a fixed interval, the magnitude of the speech spectrum, which is nonnegative, is semi-bounded, and the line spectral frequency parameters are bounded and ordered. To explore such properties of the data, some non-Gaussian statistical models, *e.g.*, the beta distribution [3], the gamma distribution [4], the Dirichlet distribution [5], were applied to describe the underlying distributions of such type of data and shown to be more efficient than the conventional Gaussian distribution based method.

The Dirichlet distribution is usually used to describe the underlying distribution of proportional data [6]. In several studies, the Dirichlet distribution based method was shown to be superior to the Gaussian distribution based method (see *e.g.*, [5]). Due to the integral expression of the gamma function and its corresponding derivatives, the MLE of the Dirichlet distribution is not analytically tractable [5]. Even though the conjugate prior of the Dirichlet distribution exists [7], the obtained posterior distribution cannot be easily used in practical problems. To derive an analytically tractable solution for Bayesian estimation, we proposed a variational inference (VI) framework [1] based method to approximately calculate the prior and posterior distributions of the Dirichlet distribution [7]. By assuming that the parameters in the Dirichlet distribution are mutually independent, a gamma distribution was assigned to each parameter (the parameters are all nonnegative) and an analytically tractable solution was obtained by using some non-linear lower-bound approximations. The proposed method works well, especially when the amount of observed data becomes larger. A similar ap-

proach was also applied in [3] for Bayesian estimation of beta distribution.

However, in the proposed VI-based method, the assumption of mutual independence violated the correlations among those parameters. When the amount of observed data is small, the shape of the parameters' joint distribution, which is informative in such case, cannot be captured efficiently by the VI-based method. To describe the correlation properly, in this paper, we assume that the joint prior distribution of all the parameters is multivariate Gaussian. By the principles of expectation propagation [1], we update each factor multivariate Gaussian distribution with the message from a observed data. The importance sampling (IS) method [1] is utilized in the updating procedure to calculate the sufficient statistics of the multivariate Gaussian distribution. Finally, a multivariate Gaussian distribution, which is the product of the entire factor multivariate Gaussian distributions, is obtained to approximate the posterior distribution of the parameters.

Unlike the VI-based method [7], the proposed EP-based method discovers the correlation among the parameters, although the nonnegativity of the parameters is violated[1]. When the amount of observed data is small, the shape of the posterior distribution, which describes the correlations among the parameters, is more important and representative than the point estimate (*e.g.*, the mean). As the number of observed data increases, the true posterior distribution will concentrate around the mode (the posterior distribution is unimodally distributed). Both the EP- and the VI-based methods can approximate the posterior distribution efficiently.

The performance of the EP-based method is evaluated and compared to the VI-based method, with several criteria. Experimental results show that the EP-based method can approximate the posterior distribution more accurate than the VI-based method, especially with smaller amount of observed data. Furthermore, the EP-based method performs more stable than the VI-based method, as it combines the messages contributed from all the observed data.

2. BAYESIAN ESTIMATION OF THE DIRICHLET DISTRIBUTION

For a vector $\mathbf{x} = [x_1, x_2, \ldots, x_K]^T$, if all the elements $x_k,\ k = 1, 2, \ldots, K$ are nonnegative and the summation of these elements equals one, the underlying distribution of this vector can be modeled by a K-dimensional Dirichlet distribution[2], which has a PDF as

$$f(\mathbf{x}) = \mathrm{Dir}(\mathbf{x};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} x_k^{\alpha_k - 1},\ \alpha_k > 0, \tag{1}$$

[1]The parameters generated from the obtained multivariate Gaussian posterior distribution have a very low probability to be negative. Also, the mean of the multivariate Gaussian posterior distribution is always positive so that we can take it as the point estimate.

[2]$\mathbf{x}$ can also be interpreted as the sample mean of a categorical multivariate distribution.

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T$ is the parameter vector contains all the free parameters and $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. To be noted, a K-dimensional Dirichlet vector variable has $K-1$ degrees of freedom. A Dirichlet distribution is unimodally distributed if all the parameters $\alpha_k,\ k = 1, 2, \ldots, K$, are greater than one. Since it is the typical case, we only study the Dirichlet distribution with all the parameters greater than one in this paper.

2.1. Conjugate Prior

As a member of the exponential family [1], the conjugate prior for the Dirichlet can be denoted as

$$f(\boldsymbol{\alpha};\boldsymbol{\beta}_0,\nu_0) = \frac{1}{C(\boldsymbol{\beta}_0,\nu_0)}\left[\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right]^{\nu_0} e^{-\boldsymbol{\beta}_0^T(\boldsymbol{\alpha}-\mathbf{1}_K)}, \tag{2}$$

where $\boldsymbol{\beta}_0 = [\beta_{1_0}, \ldots, \beta_{K_0}]^T$ and ν_0 are the hyperparameters in the prior distribution. Here, $C(\boldsymbol{\beta}_0,\nu_0)$ is the normalization factor and $\mathbf{1}_m$ denotes an m dimensional vector with all elements equal one. With Bayes' theorem and combining (1) and (2) together, we can obtain the posterior distribution of the parameters, given the observation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as[3]

$$\begin{aligned} f(\boldsymbol{\alpha}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N) =& \frac{\mathrm{Dir}(\mathbf{X}|\boldsymbol{\alpha})\, f(\boldsymbol{\alpha};\boldsymbol{\beta}_0,\nu_0)}{\int \mathrm{Dir}(\mathbf{X}|\boldsymbol{\alpha})\, f(\boldsymbol{\alpha};\boldsymbol{\beta}_0,\nu_0)\, d\boldsymbol{\alpha}} \\ =& \frac{1}{C(\boldsymbol{\beta}_N,\nu_N)}\left[\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right]^{\nu_N} e^{-\boldsymbol{\beta}_N^T(\boldsymbol{\alpha}-\mathbf{1}_K)}, \end{aligned} \tag{3}$$

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. Due to the integral expression of the gamma function, some statistics (*e.g.*, the mean vector, the covariance matrix) of the parameter $\boldsymbol{\alpha}$ cannot be obtained by analytically tractable expression. Thus, an approximation is required to approximate the prior/posterior distribution so that a closed-form expression can be derived.

2.2. Bayesian estimation with Variational Inference

Assuming that the elements in $\boldsymbol{\alpha}$ are mutually independent, a gamma prior was assigned to $\alpha_k,\ k = 1, 2, \ldots, K$ in [7]. By the principles of the VI framework [1], the prior distribution of $\boldsymbol{\alpha}$ in (2) is approximated by a product of several gamma densities as

$$f(\boldsymbol{\alpha}) \approx g(\boldsymbol{\alpha}) = \prod_{k=1}^{K}\mathrm{Gam}(\alpha_k; a_{k_0}, b_{k_0}), \tag{4}$$

where

$$\mathrm{Gam}(x;a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}. \tag{5}$$

The posterior distribution was also approximately be a product of gamma densities accordingly.

[3]To prevent any confusion, we use $f(x;a)$ to denote the PDF of x parameterized by 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.

The VI framework tries to minimize the Kullback-Leibler (KL) divergence of the true posterior distribution $f(\boldsymbol{\alpha}|\mathbf{X})$ from the approximating distribution $g(\boldsymbol{\alpha})$ as

$$\begin{aligned} g^*(\boldsymbol{\alpha}) &= \arg\min_{g(\boldsymbol{\alpha})} \{\mathrm{KL}(g \parallel f)\} \\ &= \arg\min_{g(\boldsymbol{\alpha})} \left\{ \int g(\boldsymbol{\alpha}) \ln \frac{g(\boldsymbol{\alpha})}{f(\boldsymbol{\alpha}|\mathbf{X})} d\boldsymbol{\alpha} \right\}. \end{aligned} \tag{6}$$

With non-linear lower-bound approximations, an analytically tractable solution was obtained to calculate the approximations of the posterior distribution [7]. According to [3, 7], it is reasonable to take the posterior means of the gamma densities as the point estimates of $\boldsymbol{\alpha}$. When the amount of observation is large, the point estimates become accurate.

However, the VI-based method assumed the mutually independence within the vector variable $\boldsymbol{\alpha}$, which violated the correlation among the elements in $\boldsymbol{\alpha}$. Thus, the shape of the obtained approximation is different from the true one. The difference can be neglected only when the true posterior distribution has a sharp shape and concentrates around its mode.

2.3. Bayesian estimation with Expectation Propagation

The EP is a revised version of the assumed density filtering (ADF), which is a one-pass technique for approximating the posterior distribution in Bayesian analysis [8]. The performance of ADF severely depends on the ordering of the input data. Instead of the one-pass pattern, the EP-based method goes through all the data iteratively and refines each factor distribution based on all the other factors once a time. Thus the EP-based method dose not suffer from the ordering effect of the input data.

In the EP framework, the posterior distribution of $\boldsymbol{\alpha}$ is factorized as a product of factor distributions as

$$f(\boldsymbol{\alpha}|\mathbf{X}) \propto \prod_{n=0}^{N} f_n(\boldsymbol{\alpha}), \tag{7}$$

where $f_0(\boldsymbol{\alpha})$ is the prior distribution and $f_n(\boldsymbol{\alpha}) = f(\mathbf{x}_n|\boldsymbol{\alpha})$, $n \geq 1$, is the likelihood function of $\boldsymbol{\alpha}$ given the nth observation $\mathbf{x}_n$. The EP-based method employs $g(\boldsymbol{\alpha})$, which is assumed to be a product of several factor distributions as

$$g(\boldsymbol{\alpha}) \propto \prod_{n=0}^{N} g_n(\boldsymbol{\alpha}), \tag{8}$$

to approximate the true posterior distribution. By minimizing the KL divergence of $g(\boldsymbol{\alpha})$ from $f(\boldsymbol{\alpha}|\mathbf{X})$, an optimal solution can be obtained as

$$\begin{aligned} g^*(\boldsymbol{\alpha}) &= \arg\min_{g(\boldsymbol{\alpha})} \{\mathrm{KL}(f \parallel g)\} \\ &= \arg\min_{g(\boldsymbol{\alpha})} \left\{ \int f(\boldsymbol{\alpha}|\mathbf{X}) \ln \frac{f(\boldsymbol{\alpha}|\mathbf{X})}{g(\boldsymbol{\alpha})} d\boldsymbol{\alpha} \right\}. \end{aligned} \tag{9}$$

It can be observed that the EP-based method is different from the VI-based method. These two methods minimize different forms of the KL divergence and different optimal approximations can be obtained. This difference is illustrated in Fig. 1.

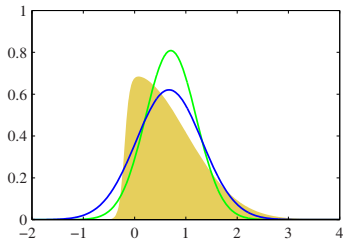


Fig. 1. Illustration of the VI- and EP-based methods. The green (narrow) curve is the VI-based approximation and the blue (broad) curve is the EP-based approximation. The yellow (shadow) shape shows the true distribution. (Figure is copied from [1, p. 508]).

The approximation obtained by the VI-based method is narrower than the true distribution while that obtained by the EP-based method is broader. It is the consequence of preventing $\ln \frac{\epsilon}{0}$ in the KL expression.

2.3.1. *EP Algorithm*

The EP framework makes a robust approximation by optimizing each factor distribution $g_n(\boldsymbol{\alpha})$ in turn with all the other factor distributions fixed. A factor distribution $g_i(\boldsymbol{\alpha})$, $i = 0, 1, \ldots, N$ is removed to get an unnormalized distribution

$$g_{\setminus i}(\boldsymbol{\alpha}) = \prod_{n=0, n\neq i}^{N} g_n(\boldsymbol{\alpha}). \tag{10}$$

Then this nunormalized distribution is combined with the ith likelihood function $f_i(\boldsymbol{\alpha})$ to get a new unnormalized distribution $\widetilde{g}(\boldsymbol{\alpha})$ as

$$\widetilde{g}(\boldsymbol{\alpha}) = g_{\setminus i}(\boldsymbol{\alpha}) f_i(\boldsymbol{\alpha}). \tag{11}$$

By the technique of moment matching, a new approximation $g^{\text{new}}(\boldsymbol{\alpha})$, which has the same distribution as $g(\boldsymbol{\alpha})$, is obtained by setting the sufficient statistics of $g^{\text{new}}(\boldsymbol{\alpha})$ equal to those of $\widetilde{g}(\boldsymbol{\alpha})$. Then the removed factor is updated as

$$g_i^*(\boldsymbol{\alpha}) \propto \frac{g^{\text{new}}(\boldsymbol{\alpha})}{g_{\setminus i}(\boldsymbol{\alpha})}. \tag{12}$$

With the above steps, the factors $g_i(\boldsymbol{\alpha})$, $i = 0, 1, \ldots, N$ will be updated iteratively till converge. Finally, the optimal approximation $g^*(\boldsymbol{\alpha})$ can be obtained by

$$g^*(\boldsymbol{\alpha}) \propto \prod_{n=0}^{N} g_n^*(\boldsymbol{\alpha}). \tag{13}$$

2.3.2. *Approximation by Multivariate Gaussian Distribution*

To (approximately) capture the correlation of $\boldsymbol{\alpha}$ in (3), we assume that each factor $g_i(\boldsymbol{\alpha})$ in (8) is multivariate Gaussian distributed as

$$\begin{aligned} g_i(\boldsymbol{\alpha}) &= \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_i, \boldsymbol{\Sigma}_i) \\ &= \frac{1}{\sqrt{2\pi}^K \sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left[-\frac{1}{2} (\boldsymbol{\alpha} - \mathbf{m}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\alpha} - \mathbf{m}_i) \right]. \end{aligned} \tag{14}$$

Then the approximation, which is a product of several multivariate Gaussian distributions, is again a multivariate Gaussian distribution as

$$g(\boldsymbol{\alpha}) = \frac{\prod_{n=0}^{N} \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_n, \boldsymbol{\Sigma}_n)}{\int \prod_{n=0}^{N} \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_n, \boldsymbol{\Sigma}_n) d\boldsymbol{\alpha}} = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}, \boldsymbol{\Sigma}), \tag{15}$$

Algorithm 1 EP-based algorithm

Input: Observation $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ from a Dirichlet distribution
Initialize: $N+1$ multivariate Gaussian factor distributions;
repeat
  for each $i = 0, 1, 2, \ldots, N$,
  refine the ith factor distribution by the methods described in (16)-(20);
until stop criterion is reached.
Calculate the mean vector and covariance matrix of the multivariate Gaussian approximation by combining $N+1$ optimal factor distributions.
Output: The optimal multivariate Gaussian approximation.

where $\mathbf{m} = \boldsymbol{\Sigma} \sum_{n=0}^{N} (\boldsymbol{\Sigma}_n^{-1} \mathbf{m}_n)$ and $\Sigma^{-1} = \sum_{n=0}^{N} \boldsymbol{\Sigma}_n^{-1}$. By removing the ith factor from $g(\boldsymbol{\alpha})$, we have

$$g_{\setminus i}(\boldsymbol{\alpha}) \propto \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_{\setminus i}, \boldsymbol{\Sigma}_{\setminus i}), \tag{16}$$

where $\mathbf{m}_{\setminus i} = \boldsymbol{\Sigma}_{\setminus i} \left(\sum_{n=0, n \neq i}^{N} \boldsymbol{\Sigma}_n^{-1} \mathbf{m}_n \right) = \boldsymbol{\Sigma}_{\setminus i} \left(\boldsymbol{\Sigma}^{-1} \mathbf{m} - \boldsymbol{\Sigma}_i^{-1} \mathbf{m}_i \right)$ and $\boldsymbol{\Sigma}_{\setminus i}^{-1} = \sum_{n=0, n \neq i}^{N} \boldsymbol{\Sigma}_n^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_i^{-1}$. The unnormalized distribution in (11) writes

$$\widetilde{g}(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_{\setminus i}, \boldsymbol{\Sigma}_{\setminus i}) \mathrm{Dir}(\mathbf{x}_i | \boldsymbol{\alpha}). \tag{17}$$

Here, we utilized the IS method [1] to get the sufficient statistics of $\widetilde{g}(\boldsymbol{\alpha})$ in (17), because $\widetilde{g}(\boldsymbol{\alpha})$ is unnormalized. To make the IS method work properly, we need to choose a reference distribution, which is easily to be sampled from and has a close shape to $\widetilde{g}(\boldsymbol{\alpha})$. To this end, we firstly apply the Laplace approximation to approximate $\widetilde{g}(\boldsymbol{\alpha})$ by a multivariate Gaussian $\breve{g}(\boldsymbol{\alpha})$ and then sample L samples $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$ from the obtained multivariate Gaussian distribution to calculate the weight of each sample as $w_l = r_l / \sum_{l=1}^{L} r_l$, $r_l = \widetilde{g}(\boldsymbol{\alpha}_l) / \breve{g}(\boldsymbol{\alpha}_l)$. Finally, the 1st and 2nd moments of $\widetilde{g}(\boldsymbol{\alpha})$ can be obtained numerically by

$$\mathbf{E}[\boldsymbol{\alpha}] \simeq \sum_{l=1}^{L} w_l \boldsymbol{\alpha}_l, \ \mathbf{E}\left[\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right] \simeq \sum_{l=1}^{L} w_l \boldsymbol{\alpha}_l \boldsymbol{\alpha}_l^T \tag{18}$$

and the mean vector and covariance matrix of the newly obtained approximation $g^{\text{new}}(\boldsymbol{\alpha})$, which is again a multivariate Gaussian distribution, are

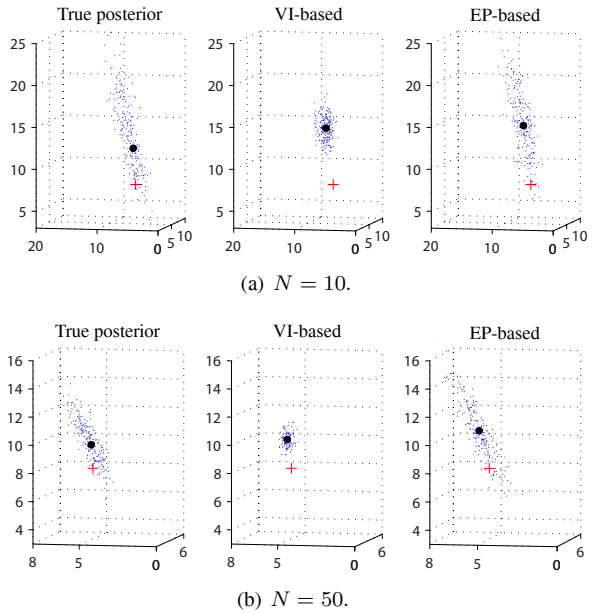
$$\mathbf{m}^{\text{new}} = \mathbf{E}[\boldsymbol{\alpha}], \ \boldsymbol{\Sigma}^{\text{new}} = \mathbf{E}\left[\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right] - \mathbf{m}^{\text{new}} (\mathbf{m}^{\text{new}})^T. \tag{19}$$

The larger L is, the more accurate the numerically calculation is. Although the Laplace approximation $\breve{g}(\boldsymbol{\alpha})$ itself is a multivariate Gaussian distribution, obtaining $g^*(\boldsymbol{\alpha})$ by IS and moment matching is more accurate than directly setting $g^*(\boldsymbol{\alpha}) = \breve{g}(\boldsymbol{\alpha})$, since the Laplace approximation captures the mode vector instead of the mean vector.

According to (12), the removed factor $g_i(\boldsymbol{\alpha})$ can now be updated by a multivariate Gaussian distribution as

$$g_i^* = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{m}_i^*, \boldsymbol{\Sigma}_i^*), \tag{20}$$

where $\mathbf{m}_i^* = \boldsymbol{\Sigma}_i^* \left[(\boldsymbol{\Sigma}^{\text{new}})^{-1} \mathbf{m}^{\text{new}} - \left(\boldsymbol{\Sigma}_{\setminus i}\right)^{-1} \mathbf{m}_{\setminus i} \right]$ and $(\boldsymbol{\Sigma}_i^*)^{-1} = (\boldsymbol{\Sigma}^*)^{-1} - \left(\boldsymbol{\Sigma}_{\setminus i}\right)^{-1}$. The above obtained covariance matrix $\boldsymbol{\Sigma}_i^*$ might be illy structured (*i.e.*, the covariance matrix is not semi-positive definite). In that case, we keep this factor distribution unchanged in this iteration and update it in the next one.



(a) $N = 10$.

(b) $N = 50$.

Fig. 2. Illustration of VI- and EP-based approximations. Data were generated from the Dirichlet distribution with $\boldsymbol{\alpha} = [3\ 5\ 8]^T$. The red cross shows the true $\boldsymbol{\alpha}$. The black dot is the mode of the corresponding distribution.

Unlike the ADF, which is a one-pass method, the EP-based algorithm will go through all the observed data and repeat the above steps to update all the factors till convergence reached. After convergence, the approximation will not be affected by the order of the input data. In principle, the EP-based method is not guaranteed to converge [1]. However, for the distribution belongs to the exponential family, if the iteration converges, the resulting solution will be a stationary point of a particular energy function. In our implementation of the EP-based algorithm for Bayesian estimation of the Dirichlet distribution, the algorithm converges in most cases. The proposed EP-based algorithm is briefly described in algorithm 1.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed EP-based algorithm was evaluated with the data generated from the Dirichlet distribution. We compare the EP-based method with the VI-based method [7], which is another method for Bayesian estimation of the Dirichlet distribution. Since these two methods minimize the KL divergences in different forms, these two methods have different properties. We will compare these two methods by both visualization and quantitative measurement.

3.1. Illustration of Different Approximations

The two-fold reasons that we choose multivariate Gaussian distribution here to approximate the posterior distribution are: 1) an analytically tractable solution can be derived for updating the factor distributions and 2) the multivariate Gaussian

distribution can (approximately) capture the correlation of the true posterior distribution.

To illustrate the differences, we generated a set of data from a given Dirichlet distribution and approximate the true posterior distribution by the EP- and VI-based methods, respectively. Fig. 2 shows an example of the comparisons. It can be observed that, in both cases ($N = 10$ and $N = 50$) the EP-based method captures the correlations among the elements in $\boldsymbol{\alpha}$ better than the VI-based method. The VI-based method, which assumes the independence among the elements in $\boldsymbol{\alpha}$, can only approximate the mode of the true posterior distribution. As the number of observations increases, the true posterior distribution, the approximation obtained by the VI, and the approximation obtained by the EP are all concentrated around the mode, thus both the VI-based method and the EP-based method can lead to reasonable point estimates, with sufficient large amount of data.

3.2. Quantitative Comparisons

For quantitative comparisons, the Absolute Evidence Difference Ratio (AEDR) was used as the criterion[4]. This quantity was evaluated with different Dirichlet distribution parameter settings and for each setting, the mean values of 50 rounds of simulations are reported.

The AEDR, which is defined as

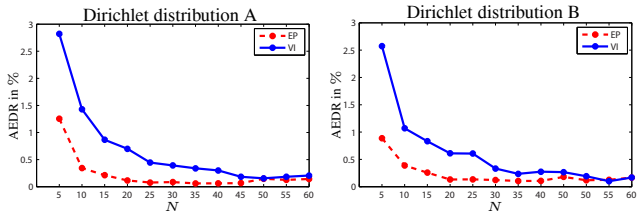
$$\text{AEDR}_m = \frac{|f_m(\mathbf{X}) - f_{\text{true}}(\mathbf{X})|}{f_{\text{true}}(\mathbf{X})} \times 100\%,\ m \in \{\text{VI}, \text{EP}\}, \qquad (21)$$

is the measurement for the relative model evidence difference between the VI-/EP-based method and the true one. The AEDR indicates how close is the approximation to the true one. Thus the model yields smaller AEDR is preferred. The model evidence $f_m(\mathbf{X})$ is calculated as

$$f_m(\mathbf{X}) = \int f(\mathbf{X}|\boldsymbol{\alpha}) g_m(\boldsymbol{\alpha}) d\boldsymbol{\alpha},\ m \in \{\text{VI}, \text{EP}\}, \qquad (22)$$

where $g_m(\boldsymbol{\alpha})$ is the posterior distribution obtained by the corresponding method. It should be noted that, in the calculation of the model evidence, we firstly estimated the posterior distribution $g_m(\boldsymbol{\alpha})$ with N samples generated from a given Dirichlet distribution. Then another 100 samples (*i.e.*, $\mathbf{X}$ in (21) and (22)) were generated from the same distribution as the upcoming observation. Thus, the posterior distribution obtained from the N samples can be considered as the "prior" distribution for the new data $\mathbf{X}$. The true model evidence was also calculated in a similar way. Here, since the integration is not analytically tractable, we utilized the rejection sampling method [1] to generate the data from the posterior distribution and then calculated the model evidence numerically. Fig. 3 shows the comparisons of the AEDRs. The EP-based method outperforms the VI-based method, with small amounts (*e.g.*, $N < 50$) of observed data. This is because that the EP-based method captured the correlation, which is informative when

[4] As the VI and the EP minimize different forms of the KL divergence, it is unfair to compare these methods by the KL divergence.



(a) $\boldsymbol{\alpha} = [3\ 5\ 8]^T$. (b) $\boldsymbol{\alpha} = [6\ 8\ 4\ 10\ 3]^T$.

Fig. 3. Comparisons of AEDRs obtained by the VI-based method and the EP-based method. Due to the limitation of space, we show only two examples here. Similar performances can also be observed for other parameter settings.

the shape of the posterior distribution is broad (the amount of the observation is small). As expected, both methods lead to efficient approximations as the amount of observation increases and the difference between these two methods cannot be distinguished. The EP-based method is also more stable (the AEDR is smaller than 1.5%) than the VI-based method.

Similar as the VI-based method, the proposed EP-based method can be extended to a mixture of Dirichlet distributions, for modeling the multimodally distributed real-life data.

4. CONCLUSION

The EP-based Bayesian estimation method for the Dirichlet distribution was proposed. The posterior distribution of the parameters in the Dirichlet distribution was approximated by a multivariate Gaussian distribution, so that the correlations among the parameters are approximately captured. Compared to the recently proposed VI-based method, the EP-based method performs better, especially when the amount of observations is small. Also, the performance of the EP-based method is more stable than that of the VI-based method.

5. REFERENCES

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

[2] Z. Ma, *Non-Gaussian Statistical Models and Their Applications*, Ph.D. thesis, KTH - Royal Institute of Technology, 2011.

[3] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, 2011.

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

[5] Z. Ma and A. Leijon, "Vector quantization of LSF parameters with a mixture of Dirichlet distributions," *IEEE Transactions on Audio, Speech, and Language Processing. To appear*, 2012.

[6] D. M. Blei, *Probabilistic models of text and images*, Ph.D. thesis, University of California, Berkeley, 2004.

[7] Z. Ma, "Variational inference of Dirichlet mixture model with extended factorzied approximation," *Submitted*, 2012.

[8] M. Opper, "A Bayesian approach to on-line learning," *On-line learning in neural networks*, pp. 363–378, 1999.