Parameter Estimation for von Mises–Fisher Mixture Model via Gaussian Distribution

Suguru Yasutomi* and Toshihisa Tanaka†
Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology
2–24–16, Nakacho, Koganei-shi, Tokyo, 184–8588, Japan
E-mail: *yasutomi@sip.tuat.ac.jp, †tanakat@cc.tuat.ac.jp

***Abstract*—Directional statistics deal with direction, such as angles and phases. A well-known distribution in directional statistics is von Mises–Fisher (vMF) distribution, which is Gaussian distribution on a unit hypersphere. For a vMF mixture model, a maximum likelihood estimator and a variational Bayes estimator have already been derived. However, an iterative algorithm for finding the maximum likelihood estimator may accumulate approximation error. Besides, the variational Bayes estimator cannot estimate some parameters. This paper derives an estimator of the parameters in the vMF mixture model via the Gaussian distribution to solve these problems. We focus on the fact that the vMF distribution is derived from the Gaussian distribution. At first, we apply the estimation for the Gaussian mixture model to observed samples. Then, we convert the estimated Gaussian mixture distribution to a vMF mixture distribution. Experimental results support the analysis.**

I. Introduction

Random variables observed as directions, such as angles and phases can be dealt with by directional statistics [1]. In directional statistics, the direction is the random variable. For univariate directional data, the random variable would be a circular variable. In that case, the von Mises distribution, which is a circular Gaussian distribution, is popularly used [1]. For multivariate directional data, it can be expressed as random vectors that have unit norm. Thus, these data are on a unit hypersphere. The von Mises–Fisher (vMF) distribution, which is a natural extension of the Gaussian distribution on the unit hypersphere, is primarily used in directional statistics [1]. The vMF distribution has two parameters called the mean, and the concentration. These parameters correspond to the mean and the precision in Gaussian distribution, respectively. The vMF distribution has a wide range of applications such as text mining, document clustering [2], [3], wind direction analysis [4], protein clustering [5] and tractography [6].

Parameter estimation of the distribution from the observed data is fundamental problem in engineering. Well-known solutions to the problem are the maximum likelihood estimation and the Bayesian estimation. In the maximum likelihood estimation, parameters are estimated by maximizing the likelihood function. For a mixture model, an iterative algorithm called Expectation Maximization (EM) algorithm [7] is efficiently applied [8]. The Bayesian estimation approach estimates the posterior distributions of the parameters. The point estimations are given by the expectations of the posterior distributions. However, the posterior distributions cannot be obtained in the closed form for some practical models [8]. In such cases, approximation of posterior distributions such as the variational Bayes (VB) can be used.

Banejee et al. [9] estimated the parameters of the vMF mixture models with EM algorithm. In the iterative algorithm, the inverse function of the ratio of the Bessel function is necessary which has no closed form. Therefore, approximation of it is needed, and the approximation error may be accumulated during the iteration. Tanabe et al. [10] derived the VB inference for vMF mixture models, but it cannot estimate some parameters. For von Mises mixture models, the two dimensional case of the vMF, Tanaka and Kobayashi [11] derived the VB inference via the bivariate Gaussian distribution.

This paper proposes a method for estimating the parameters in the vMF mixture models via Gaussian distribution extending [11]. We focus on the fact that the vMF distribution is derived from the Gaussian distribution. The underlying idea behind the proposed method is that we first apply an estimation method for the Gaussian mixture model, and then convert the estimated Gaussian mixture distribution into a vMF mixture distribution. The proposed estimator offers the iterative algorithm without the above mentioned approximation for the EM algorithm, and can estimate all the parameters in the VB. Experimental results supports the analysis.

II. Parameter Estimation of vMF Mixture Model

A. von Mises–Fisher (vMF) Distribution

The vMF distribution [1] is a natural extension of the Gaussian distribution on a unit hypersphere. The p-dimensional vMF distribution has the density function defined as

$$\mathrm{vMF}_p(\mathbf{x}|\boldsymbol{\mu},\kappa) = C_p(\kappa)\exp[\kappa\boldsymbol{\mu}^\intercal\mathbf{x}],\ \|\mathbf{x}\| = 1, \tag{1}$$

where $\boldsymbol{\mu}$ is a p-dimensional unit vector called mean direction, κ is the concentration parameter, and $C_p(\kappa)$ is the normalization term defined as

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}I_{p/2-1}(\kappa)}, \tag{2}$$

where $I_\nu(x)$ represents the modified Bessel function of the first kind of order ν that defined as

$$I_\nu(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{x\cos t}\cos(\nu t)\mathrm{d}t. \tag{3}$$

Fig. 1 shows an example of samples which have the vMF dis-

978-616-361-823-8 © 2014 APSIPA

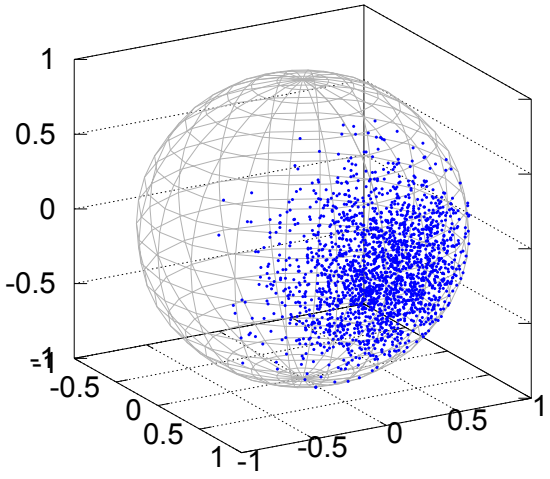

Fig. 1. Example of samples distributed from vMF distribution where $p = 3, \boldsymbol{\mu} = [1\ 0\ 0]^\intercal, \kappa = 10$.

tribution. The expectation of the vMF distribution is obtained as

$$\mathbb{E}[\mathbf{x}] = A_p(\kappa)\boldsymbol{\mu} = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}\boldsymbol{\mu}. \tag{4}$$

From (1), vMF mixture distribution can be defined as

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \sum_{k=1}^{K} \pi_k \mathrm{vMF}_p(\mathbf{x}|\boldsymbol{\mu}_k, \kappa_k), \tag{5}$$

where K is the number of the mixture components, and π_k is the mixing weight of the k-th component.

B. Deriving vMF from Gaussian

We show that the vMF distribution can be derived from the Gaussian distribution. Consider a Gaussian distribution with the covariance matrix that is multiplication of identity matrix like $\sigma^2\mathbf{I}$. The distribution is expressed as follows:

$$\mathcal{N}_p(\mathbf{x}|\mathbf{m}, \sigma^2\mathbf{I}) = \frac{1}{(2\pi)^{p/2}}\frac{1}{\sigma^p}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{m})^\intercal(\mathbf{x}-\mathbf{m})\right]. \tag{6}$$

This is called an isotropic Gaussian distribution. A vMF distribution can be derived by projecting the isotropic Gaussian distribution onto a unit hypersphere. More specifically, we can obtain a vMF distribution by applying the following transformations to an isotropic Gaussian distribution:

$$\|\mathbf{x}\| = 1, \tag{7}$$
$$\mathbf{m} = r_0\boldsymbol{\mu}, \tag{8}$$
$$\sigma^2 = \frac{r_0}{\kappa}. \tag{9}$$

C. vMF Mixture Model

The likelihood function of the vMF mixture model is given as

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathrm{vMF}_p(\mathbf{x}|\boldsymbol{\mu}_k, \kappa_k)\right\}. \tag{10}$$

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$ having a vMF mixture distribution.The parameters to be estimated are the mixing weights π_k, mean directions $\boldsymbol{\mu}_k$, and concentration parameters κ_k.

D. EM Algorithm for vMF Mixture Model

We review the EM algorithm for the vMF mixture model derived by Banarjee et al. [9]. Let $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ be a set of latent variables. This latent variable denotes from which component a sample comes. Each latent variable, $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, corresponds to each observed data, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, respectively. Specifically $\mathbf{z}_n$ is a K-dimensional vector that has a 1-of-K representation described as

$$\sum_{k=1}^{K} z_{nk} = 1, \ z_{nk} \in 0, 1. \tag{11}$$

Maximum likelihood estimators of parameter $\boldsymbol{\mu}_k, \kappa_k$ and π_k are given by maximizing (10). In the EM framework, we find following equations:

$$\boldsymbol{\mu}_k = \frac{\mathbf{r}}{\|\mathbf{r}\|}, \tag{12}$$

$$A_p(\kappa_k) = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mu_k^\intercal\mathbf{x}_n, \tag{13}$$

$$\pi_k = \frac{N_k}{N}, \tag{14}$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathrm{vMF}_p(\mathbf{x}_n|\boldsymbol{\mu}_k, \kappa_k)}{\sum_{j=1}^{K}\pi_j \mathrm{vMF}_p(\mathbf{x}_n|\boldsymbol{\mu}_j, \kappa_j)}, \tag{15}$$

where

$$\mathbf{r} = \sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n, \tag{16}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}), \tag{17}$$

$$A_p(x) = \frac{I_{p/2}(x)}{I_{p/2-1}(x)}. \tag{18}$$

In (15), $\gamma(z_{nk})$, which is called responsibility is a probability that the latent variable z_{nk} becomes 1.

Calculating (12) to (14) and (15) iteratively gives the maximum likelihood solution. Eq. (15) is called the E-step that calculates the responsibility using present parameters. Eqs. (12) to (14) are called the M-step that updates parameters based on the responsibility. These two steps are repeated until the parameters converge.

In (13), concentration parameters, κ_k, are implicitly given in $A_p(\kappa_k)$, which is the ratio of modified Bessel function of the first kind as described in (13). However, it is not possible to obtain a closed form expression for $A_p^{-1}(x)$. Therefore, the approximation of $A_p^{-1}(x)$ is necessary [9], [12], [13]. In [9], an approximation of $A_p^{-1}(x)$ is given as

$$A_p^{-1}(x) \approx \frac{px - x^3}{1 - x^2}. \tag{19}$$

It should be noted that this EM approach may accumulate the approximation error because of the approximation of $A_p(\kappa_k)$.

E. VB Inference for vMF Mixture Model

We review the VB inference for vMF mixture model derived by Tanabe et al. [10] and show that it cannot estimate the concentration parameters. Let $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ be a set of latent variables defined in the same way as the EM approach. Given the mixing weights $\boldsymbol{\pi} = \{\pi_k\}$, the conditional distribution of $\mathbf{Z}$ can be expressed as

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}. \tag{20}$$

Besides, the conditional distribution of $\mathbf{X}$ is given as

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathrm{vMF}_p(\mathbf{x}_n|\boldsymbol{\mu}_k, \kappa_k)^{z_{nk}}. \tag{21}$$

As a prior distribution for the mixing weights in the same manner as the VB for Gaussian mixture model [8] the Dirichlet distribution is used. The Dirichlet distribution is defined as

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}. \tag{22}$$

For the mean $\boldsymbol{\mu}$ and the concentration parameter $\boldsymbol{\kappa}$, the vMF-Gamma distribution given by

$$p(\boldsymbol{\mu}, \boldsymbol{\kappa}) = \prod_{k=1}^{K} \mathrm{vMF}_p(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \beta_0\kappa_k)\mathrm{Gam}(\kappa_k|a_0, b_0) \tag{23}$$

is used as a prior distribution.

In the VB approach, a posterior distribution is assumed to be approximated as

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = q(\mathbf{Z})q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\kappa}). \tag{24}$$

Parameters are found so as to minimize a KL divergence between the true posterior distribution and the distribution q. Such $q^\star$ is given [8] as

$$\ln q^\star(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\kappa}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa})] + \mathrm{const}, \tag{25}$$

$$\ln q^\star(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\kappa}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa})] + \mathrm{const}, \tag{26}$$

$$\ln q^\star(\boldsymbol{\mu}, \boldsymbol{\kappa}) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa})] + \mathrm{const}, \tag{27}$$

where $q^\star$ is called variational distribution. The joint distribution is assumed to be

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}). \tag{28}$$

Based on the above preparation, the VB for the vMF mixture model derived as Appendix A. In the resulting algorithm [10], the concentration parameters κ_k are assumed known. This limitation comes from the derivation of the algorithm as seen below.

Substituting (28) for (25), $q^\star(\mathbf{Z})$ can be given as

$$\ln q^\star(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + \mathrm{const}, \tag{29}$$

where const is a constant independent of $\mathbf{Z}$, and

$$\ln\rho_{nk} = \mathbb{E}[\ln\pi_k] + \left(\frac{p}{2} - 1\right)\mathbb{E}[\ln\kappa_k] - \frac{p}{2}\ln(2\pi), \\ - \mathbb{E}[\ln I_{p/2-1}(\kappa_k)] + \mathbb{E}_{\boldsymbol{\mu}_k,\kappa_k}[\kappa_k\boldsymbol{\mu}_k^\mathsf{T}\mathbf{x}_n]. \tag{30}$$

The fourth term on the right-hand side of (30) is the expectation of the modified Bessel function of the first kind that is not given in a closed form. This implies that this algorithm cannot estimate the concentration parameters.

III. Proposed Estimation of vMF Parameters

A. Parameter Estimation via Gaussian Distribution

Under the assumption that $\mathbf{x}_n$ is from the k-th component, (4) with respect to the k-th component is given as

$$\mathbb{E}[\mathbf{x}_n] = A_p(\kappa_k)\boldsymbol{\mu}_k. \tag{31}$$

However, we cannot calculate $\mathbb{E}[\mathbf{x}_n]$. To cope with this, we propose to replace $\mathbb{E}[\mathbf{x}_n]$ by $\hat{\mathbf{m}}_k$, which is the estimation of the mean of the isotropic Gaussian mixture model. Then, from (31), we get

$$\boldsymbol{\mu}_k \simeq \frac{\hat{\mathbf{m}}_k}{\|\hat{\mathbf{m}}_k\|}, \tag{32}$$

$$\kappa_k \simeq A_p^{-1}(\|\hat{\mathbf{m}}_k\|). \tag{33}$$

It should be noted that (32) and (33) give transformations from an isotropic Gaussian distribution to a vMF distribution. The validity of the transformations is discussed in the next section. In (33), $A_p^{-1}(\cdot)$ will be approximated by using (19), in the same way as [9].

We focus on the relationship between the vMF distribution and the Gaussian distribution to derive parameter estimation method of the vMF mixture model. At first, the parameter estimation methods for isotropic Gaussian mixture model are applied to observed data. For the Gaussian mixture model, EM algorithm and VB inference are well-known [8]. Next, we convert estimated parameters of the isotropic Gaussian mixture distribution into the ones of a vMF mixture distribution. The underlying idea behind the proposed method is that for the case of Gaussian mixture models parameter estimation can be achieved by well-established algorithms such as the EM algorithm and the VB inference.

B. Validity of Derived Transformations

To validate that the replacement of the expectation $\mathbb{E}[\mathbf{x}_n]$ by the estimator $\hat{\mathbf{m}}_k$ in (31) is proper, we show the consistency [14], [15] of the estimator $\hat{\mathbf{m}}_k$. The consistency is verified by confirming that

$$\mathbb{E}_{\mathbf{x}}\left\|\hat{\mathbf{m}}_k - \mathbb{E}\left[\mathbf{x}_n\right]\right\|^2 = \mathbb{E}_{\mathbf{x}}\left\|\hat{\mathbf{m}}_k - A_p(\kappa_k)\boldsymbol{\mu}\right\|^2 \tag{34}$$

becomes 0 when number of samples goes to infinity ($N \to \infty$).

First, we consider the case that parameters of the isotropic Gaussian mixture model are estimated with the maximum

likelihood method. In the non-mixed case, i.e. the number of components $K = 1$, with (4) and (34), we find

$$\mathbb{E}_{\mathbf{x}} \left\| \hat{\mathbf{m}} - A_p(\kappa)\boldsymbol{\mu} \right\|^2 = \mathbb{E}_{\mathbf{x}} \left\| \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n - A_p(\kappa)\boldsymbol{\mu} \right\|^2 = \frac{1}{N} - \frac{1}{N} A_p^2(\kappa) \to 0 \ (N \to \infty). \tag{35}$$

This verifies the consistency. In the K-mixture case, estimator of mean is expressed using the responsibility $\gamma(z_{nk})$ as

$$\hat{\mathbf{m}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n, \tag{36}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{37}$$

Suppose that latent variable $\mathbf{z}_n$ is known, or responsibility $\gamma(z_{nk})$ is estimated properly. Then (34) with respect to the k-th mixture component can be written as

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}} \left\| \hat{\mathbf{m}}_k - A_p(\kappa_k)\boldsymbol{\mu}_k \right\|^2 \\ &= \mathbb{E}_{\mathbf{x}}[\mathbf{m}_k^\intercal \mathbf{m}_k] - 2A_p(\kappa_k)\mathbb{E}_{\mathbf{x}}[\boldsymbol{\mu}_k^\intercal \mathbf{m}_k] + A_p^2(\kappa_k) \\ &= \frac{1}{N_k} \left\{ \sum_{n=1}^{N} \gamma^2(z_{nk}) + A_p(\kappa_k) \left(N_k^2 - \sum_{n=1}^{N} \gamma^2(z_{nk}) \right) \right\} \\ &\qquad - 2A_p^2(\kappa_k) + A_p^2(\kappa_k) \\ &= \frac{\sum_{n=1}^{N} \gamma^2(z_{nk})}{N_k^2} \left(1 - A_p^2(\kappa)\right) \to 0 \ (N \to \infty). \end{aligned} \tag{38}$$

This verifies the consistency of K-mixture case.

Next, we consider about VB. In VB, when $N \to \infty$, or when the prior distribution is a uninformative prior, the mean estimator is given by (36), same as the EM approach [8]. Thus, in such case, the consistency can be regarded as valid even if the VB method applied instead of the EM method.

IV. Experiments and Results

We evaluate the proposed parameter estimation method with simulated data. We generated the samples that have the vMF mixture distribution by the method described in [16]. We applied the following parameter estimation methods to the data for the comparison of the performance. The methods we compared were the proposed EM approach, the proposed VB approach, the conventional EM approach [9], and the conventional VB approach [10]. Table I shows eight different vMF mixture models we setup. In the column of $\boldsymbol{\mu}_k$ in Table I, "random" means that the means are set uniformly at random on the p-dimensional unit hypersphere satisfying $\boldsymbol{\mu}_i^\intercal \boldsymbol{\mu}_j < 0.25$, where $i, j \in \{1, \ldots, K\}$, $i \neq j$. In all the experiments, we randomly initialized parameters. It is known that the EM approach and the VB approach may yield improper estimate of parameters depending on the initial parameters. Thus, we consider that the estimation has failed when $\varepsilon(\pi_k) = \left| \frac{\pi_k - \hat{\pi}_k}{\pi_k} \right| > 1$ and applied the estimation again, reinitializing the parameters. In the VB approach, we set hyperparameters of the prior distributions in an uninformative way.

TABLE I
Parameters of vMF mixture models

	N	p	K	k	π_k	$\boldsymbol{\mu}_k$	κ_k
Model 1	1000	3	1	1	1	$[1\ 0\ 0]^\intercal$	5
Model 2	100	3	1	1	1	$[1\ 0\ 0]^\intercal$	5
Model 3	1000	20	1	1	1	$[1\ 0\ \cdots\ 0]^\intercal$	10
Model 4	100	20	1	1	1	$[1\ 0\ \cdots\ 0]^\intercal$	10
Model 5	1000	3	2	1	0.4	random	10
				2	0.6	random	5
Model 6	2000	3	3	1	0.3	random	20
				2	0.4	random	25
				3	0.3	random	30
Model 7	3000	3	5	1	0.2	random	22
				2	0.2	random	24
				3	0.2	random	26
				4	0.2	random	28
				5	0.2	random	30
Model 8	2000	5	3	1	0.3	random	20
				2	0.4	random	25
				3	0.3	random	30

Table II shows the results of the experiments. We conducted data generation 600 times with each model, and estimated parameters by applying each methods. In models 5–8, the means are set randomly before each data generation. In Table II, bold-faced numbers indicate the better performance compared to the conventional and proposed methods. Some of the standard deviations in Table II are zero since the estimated values are almost the same through 600 trials. It can be seen that all the methods in non-mixture models 1–4 showed the same performance in estimation. In models 5–7 which have multiple mixture components and are 3-dimensional models, and in model 8 which has higher dimensions than models 5–7, the proposed methods achieved estimation precision although competitive to the conventional methods. The proposed variational Bayes approach succeeded to estimate concentration parameters accurately, the conventional one cannot estimate it.

V. Conclusions

In this paper, we derived a parameter estimation method of von Mises–Fisher mixture model through Gaussian distributions. The experimental results show that the proposed method achieves competitive accuracy in parameter estimation to conventional methods. Furthermore, the proposed VB method can estimate concentration parameters that cannot be estimated with the conventional method.

The proposed method assumes that number of mixture components is known. Besides, the estimation result depend on the initial parameters. Thus, estimating the number of components and reducing the dependency on the initial parameters would be a future work.

TABLE II

EXPERIMENTAL RESULTS. $\varepsilon(\pi_k) = \text{avg}\left|\frac{\pi_k - \hat{\pi}_k}{\pi_k}\right|, c(\boldsymbol{\mu}_k) = \text{avg}\left(\boldsymbol{\mu}_k^\intercal \hat{\boldsymbol{\mu}}_k\right), \varepsilon(\kappa_k) = \text{avg}\left|\frac{\kappa_k - \hat{\kappa}_k}{\kappa_k}\right|$. THE NUMBERS IN PARENTHESES ARE THE STANDARD DEVIATIONS.

	Conventional (EM)[9]			Conventional (VB)[10]			Proposed (EM)			Proposed (VB)		
	$\varepsilon(\pi_k)$	$c(\boldsymbol{\mu}_k)$	$\varepsilon(\kappa_k)$	$\varepsilon(\pi_k)$	$c(\boldsymbol{\mu}_k)$	$\varepsilon(\kappa_k)$	$\varepsilon(\pi_k)$	$c(\boldsymbol{\mu}_k)$	$\varepsilon(\kappa_k)$	$\varepsilon(\pi_k)$	$c(\boldsymbol{\mu}_k)$	$\varepsilon(\kappa_k)$
Model 1	0.000 (±0.000)	1.000 (±0.000)	0.053 (±0.031)	0.000 (±0.000)	1.000 (±0.000)	-	0.000 (±0.000)	1.000 (±0.000)	0.053 (±0.031)	0.000 (±0.000)	1.000 (±0.000)	0.053 (±0.031)
Model 2	0.000 (±0.000)	0.998 (±0.003)	0.100 (±0.079)	0.000 (±0.000)	0.998 (±0.003)	-	0.000 (±0.000)	0.998 (±0.003)	0.100 (±0.079)	0.000 (±0.000)	0.998 (±0.003)	0.100 (±0.079)
Model 3	0.000 (±0.000)	0.998 (±0.001)	0.015 (±0.012)	0.000 (±0.000)	0.998 (±0.001)	-	0.000 (±0.000)	0.998 (±0.001)	0.015 (±0.012)	0.000 (±0.000)	0.998 (±0.001)	0.015 (±0.012)
Model 4	0.000 (±0.000)	0.978 (±0.007)	0.058 (±0.044)	0.000 (±0.000)	0.978 (±0.007)	-	0.000 (±0.000)	0.978 (±0.007)	0.058 (±0.044)	0.000 (±0.000)	0.978 (±0.007)	0.058 (±0.044)
Model 5	0.008 (±0.009)	1.000 (±0.000)	**0.060** (±0.045)	0.092 (±0.152)	0.993 (±0.019)	-	0.008 (±0.010)	1.000 (±0.000)	0.069 (±0.050)	**0.044** (±0.047)	**0.999** (±0.002)	0.104 (±0.070)
Model 6	**0.012** (±0.086)	**0.989** (±0.125)	**0.044** (±0.075)	**0.002** (±0.004)	**1.000** (±0.000)	-	0.015 (±0.096)	0.986 (±0.142)	0.049 (±0.129)	0.003 (±0.042)	0.997 (±0.065)	0.038 (±0.045)
Model 7	0.001 (±0.001)	1.000 (±0.000)	0.037 (±0.028)	0.003 (±0.003)	1.000 (±0.000)	-	0.001 (±0.001)	1.000 (±0.000)	0.037 (±0.028)	**0.001** (±0.001)	1.000 (±0.000)	0.037 (±0.028)
Model 8	0.026 (±0.126)	0.977 (±0.173)	0.042 (±0.094)	0.004 (±0.007)	1.000 (±0.000)	-	**0.021** (±0.117)	**0.981** (±0.154)	**0.040** (±0.089)	**0.001** (±0.002)	1.000 (±0.000)	0.028 (±0.021)

APPENDIX

A. Derivation of VB Inference for vMF Mixture Model

We introduce VB inference, which we reviewed in Section II-E, derived by Tanabe et al. [10]. Assuming that the concentration parameters are known, parameters to estimate are the mean and mixing weight of each mixture component.

For the prior distribution of the mean $\boldsymbol{\mu}$, the vMF distribution given by

$$p(\boldsymbol{\mu}) = \prod_{k=1}^{K} \text{vMF}_p(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, \kappa_0) \tag{39}$$

is used instead of (23). In [10], κ_0 is assumed to be 0 so that (39) be an uninformative prior. The approximated posterior distribution that given by (24) is rewritten as

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}) = q(\mathbf{Z}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}). \tag{40}$$

Considering the assumption above, we can rewrite (25) to (27) as follows:

$$\ln q^\star(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu})] + \text{const}, \tag{41}$$

$$\ln q^\star(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\mu}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu})] + \text{const}, \tag{42}$$

$$\ln q^\star(\boldsymbol{\mu}) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu})] + \text{const}. \tag{43}$$

The joint distribution is assumed to be

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}) p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}). \tag{44}$$

For $q^\star(\mathbf{Z})$, substituting (44) to (41), we obtain

$$\begin{aligned}
\ln q^\star(\mathbf{Z}) =& \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}) + \ln p(\mathbf{Z}|\boldsymbol{\pi})] + \text{const} \\
=& \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} \ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa})^{z_{nk}} \right. \\
& \left. + \sum_{n=1}^{N}\sum_{k=1}^{K} \ln \pi_k^{z_{nk}}\right] + \text{const} \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const},
\end{aligned} \tag{45}$$

where const is a constant independent of $\mathbf{Z}$, and

$$\begin{aligned}
\ln \rho_{nk} =& \mathbb{E}[\ln \pi_k] + \left(\frac{p}{2} - 1\right) \ln \kappa_k - \frac{p}{2} \ln(2\pi) \\
& - \ln I_{p/2-1}(\kappa_k) + \kappa_k \mathbb{E}_{\boldsymbol{\mu}_k}[\boldsymbol{\mu}_k^\intercal \mathbf{x}_n].
\end{aligned} \tag{46}$$

From (45), $q^\star(\mathbf{Z})$ can be expressed as

$$q^\star(\mathbf{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}. \tag{47}$$

Normalizing (47), we obtain

$$q^\star(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}, \tag{48}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}. \tag{49}$$

To calculate $q^\star(\mathbf{Z})$, $q^\star(\boldsymbol{\pi})$ and $q^\star(\boldsymbol{\mu})$ are necessary.

For $q^\star(\boldsymbol{\pi})$, substituting (44) to (42), we obtain

$$\begin{aligned}
\ln q^\star(\boldsymbol{\pi}) =& \mathbb{E}_{\mathbf{Z},\boldsymbol{\mu}}[\ln p(\mathbf{Z}|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi})] + \text{const} \\
=& \mathbb{E}_{\mathbf{Z},\boldsymbol{\mu}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \pi_k + \sum_{k=1}^{K} \ln \pi_k^{\alpha_0 - 1}\right] + \text{const} \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k^{\alpha_0 - 1} + \text{const} \\
=& \sum_{n=1}^{N} N_k \ln \pi_k + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k^{\alpha_0 - 1} + \text{const},
\end{aligned} \tag{50}$$

where const is a constant independent of $\boldsymbol{\pi}$, and

$$\mathbb{E}[z_{nk}] = r_{nk}, \tag{51}$$

$$\sum_{k=1}^{K} r_{nk} = N_k. \tag{52}$$

Taking exponential both sides of (50), it can be expressed as a Dirichret distribution given as

$$q^\star(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}), \tag{53}$$

where $\boldsymbol{\alpha} = \{\alpha_0 + N_k\}$.

For $q^\star(\boldsymbol{\mu})$, substituting (44) to (43), we obtain

$$\begin{aligned}
\ln q^\star(\boldsymbol{\mu}) =& \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\kappa}) + \ln p(\boldsymbol{\mu})] + \mathrm{const} \\
=& \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \mathrm{vMF}(\mathbf{x}_n|\boldsymbol{\mu}_k,\kappa_k) \right. \\
& \left. + \sum_{k=1}^{K} \ln \mathrm{vMF}(\mathbf{x}_n|\boldsymbol{\mu}_k, 0)\right] + \mathrm{const} \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[z_{nk}(\ln C(\kappa_k) + \kappa_k \boldsymbol{\mu}_k^\intercal \mathbf{x}_n)] + \mathrm{const},
\end{aligned} \tag{54}$$

where const is a constant independent of $\boldsymbol{\mu}$. From (54), it is clear that $q^\star(\boldsymbol{\mu})$ can be expanded as

$$q^\star(\boldsymbol{\mu}) = \prod_{k=1}^{K} q^\star(\boldsymbol{\mu}_k). \tag{55}$$

Thus, we consider about $q^\star(\boldsymbol{\mu}_k)$. From (54), $q^\star(\boldsymbol{\mu}_k)$ is written as

$$\begin{aligned}
\ln q^\star(\boldsymbol{\mu}) =& \sum_{n=1}^{N} \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[z_{nk}(\ln C(\kappa_k) + \kappa_k \boldsymbol{\mu}_k^\intercal \mathbf{x}_n)] + \mathrm{const} \\
=& \sum_{n=1}^{N} r_{nk}\kappa_k \boldsymbol{\mu}_k^\intercal \mathbf{x}_n + \mathrm{const} \\
=& N_k \kappa_k \boldsymbol{\mu}_k^\intercal \bar{\mathbf{x}}_k + \mathrm{const},
\end{aligned} \tag{56}$$

where

$$\bar{\mathbf{x}}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n. \tag{57}$$

Taking exponential both sides of (56), it can be expressed as a vMF distribution given as

$$q^\star(\boldsymbol{\mu}) = \mathrm{vMF}(\boldsymbol{\mu}_k|\bar{\mathbf{x}}, N_k\kappa_k). \tag{58}$$

Using (53) and (58), the expectations in (46) are given as

$$\mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}), \tag{59}$$

$$\mathbb{E}[\boldsymbol{\mu}_k^\intercal \mathbf{x}_n] = A_p(N_k\kappa_k)\bar{\mathbf{x}}_k^\intercal \mathbf{x}_n, \tag{60}$$

where $\psi(\cdot)$ denotes the digamma function and $\hat{\alpha} = \sum_{k=1}^{K} \alpha_k$. In this VB inference method, approximated posteriors are obtained by calculating (48) (VB E-step), and (53) and (58) (VB M-step) iteratively.

REFERENCES

[1] K. V. Mardia and P. E. Jupp, *Directional Statistics*, ser. Wiley series in probability and statistics. Wiley, 2000.

[2] N. K. Anh, N. T. Tam, and N. Van Linh, "Document clustering using dirichlet process mixture model of von Mises–Fisher distributions," in *Proceedings of the Fourth Symposium on Information and Communication Technology*, ser. SoICT '13. New York, NY, USA: ACM, 2013, pp. 131–138.

[3] S. Gopal and Y. Yang, "Von Mises–Fisher clustering models," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 154–162, to be published.

[4] J. A. Carta, C. Bueno, and P. Ramírez, "Statistical modelling of directional wind speeds using mixtures of von mises distributions: Case study," *Energy Conversion and Management*, vol. 49, no. 5, pp. 897–907, May 2008.

[5] P. J. Green and K. V. Mardia, "Bayesian alignment using hierarchical models, with applications in protein bioinformatics," *Biometrika*, vol. 93, no. 2, pp. 235–254, 2006.

[6] F. Zhang, E. R. Hancock, C. Goodlett, and G. Gerig, "Probabilistic white matter fiber tracking using particle filtering and von Mises–Fisher sampling," *Medical Image Analysis*, vol. 13, no. 1, pp. 5–18, Feb. 2009.

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

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

[9] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway, "Clustering on the unit hypersphere using von Mises–Fisher distributions," *Journal of Machine Learning Research*, vol. 6, no. 9, pp. 1345–1382, 2005.

[10] A. Tanabe, K. Fukumizu, S. Oba, T. Takenouchi, and S. Ishii, "A variational bayes inference for a mixture of von Mises–Fisher distribution," *Technical Report of IEICE*, vol. 104, no. 758, pp. 131–136, Mar. 2005.

[11] T. Tanaka and M. Kobayashi, "Variational bayes for directional statistics," *Technical Report of IEICE*, vol. 113, no. 120, pp. 113–118, Jul. 2013.

[12] A. Tanabe, K. Fukumizu, S. Oba, T. Takenouchi, and S. Ishii, "Parameter estimation for von Mises–Fisher distributions," *Computational Statistics*, vol. 22, no. 1, pp. 145–157, Mar. 2007.

[13] S. Sra, "A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$," *Computational Statistics*, vol. 27, no. 1, pp. 177–190, Feb. 2011.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[15] R. Hogg, J. McKean, and A. Craig, *Introduction to mathematical statistics*, ser. Pearson education international. Pearson Education, 2005.

[16] A. T. A. Wood, "Simulation of the von Mises Fisher distribution," *Communications in Statistics - Simulation and Computation*, vol. 23, no. 1, pp. 157–164, 1994.