Approximate variational inference in Dirichlet process von Mises–Fisher mixture models

Ulpu Remes, Shreyas Seshadri, Okko Räsänen

October 24, 2016

1 Introduction

This document describes the approximate variational inference method that we have used to estimate von Mises–Fisher (VMF) mixture models with Dirichlet process (DP) priors. The method is based on the variational methods introduced in [1, 2].

2 Model

The observations $\boldsymbol{x}(n)$ considered in this work are modelled as *i.i.d* samples from an infinite mixture model constructed as follows. We introduce a latent indicator variable $z(n)$ to indicate the mixture component $k = 1 \dots \infty$ that produced observation n and write the observations probabilities

$$p(\boldsymbol{x}(n)|z(n), \boldsymbol{\phi}) = \prod_{k=1}^{\infty} p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))^{\mathbf{1}[z(n)=k]}, \tag{1}$$

where $\boldsymbol{\phi}(k)$ denote the distribution parameters associated with mixture component k and $\mathbf{1}[z(n) = k]$ is an indicator function that evaluates to 1 if $z(n) = k$ and 0 otherwise. We assume that the allocations $z(n)$ follow a categorical distribution $\boldsymbol{\pi}$ constructed based on the stick-breaking process. The allocation probabilities can be expressed as

$$p(z(n) = k|\boldsymbol{\pi}) = \prod_{k=1}^{\infty} (\pi(k))^{\mathbf{1}[z(n)=k]}, \tag{2}$$

where $\pi(k)$ denote event probabilities. We assume that the event probabilities are constructed by successively breaking a unit stick into infinite number of pieces. The event probabilities are determined as

$$\pi(k) = v(k) \prod_{i=1}^{k-1} (1 - v(i)), \tag{3}$$

where $v(k)$ denote stick proportions. Equation (3) can be substituted in Equation (2) and the dependencies between cluster allocations and stick proportions

can be expressed as

$$p(z(n) = k|\boldsymbol{v}) = \prod_{k=1}^{\infty} (1 - v(k))^{\mathbf{1}[z(n)>k]} v(k)^{\mathbf{1}[z(n)=k]}. \tag{4}$$

The stick proportions are modelled as $i.i.d$ random variables with prior distribution $Beta(1, \alpha)$, which means that the parameter vector $\boldsymbol{\phi}^*(n)$ associated with observation n can be understood as a random variable with prior distribution $DP(\alpha, H)$. Moreover the complete model can be understood as a DP that has been extended with the observation model $p(\boldsymbol{x}(n)|\boldsymbol{\phi}^*(n))$ (Equation 1). We assume that the concentration parameter α and base distribution H are modelled as hyperparameters. The concentration parameter can also be modelled as a random variable as proposed in [1] or prior distributions other than $Beta(1, \alpha)$ can be used with the stick proportions (Equation 3) to construct models discussed in [3].

The current work focusses on directional data and component-condition observation likelihoods $p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))$ modelled as von Mises–Fisher (VMF) distributions [4]. Thus we assume that the observations $\boldsymbol{x}(n)$ are D-dimensional unit vectors, $||\boldsymbol{x}(n)|| = 1$. The observations can be understood as points on $D-1$-dimensional unit hypersphere, and observed clusters can be characterised based on the mean direction and concentration around the mean. The component-conditioned likelihoods are calculated as

$$p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k)) = \frac{1}{Z(\lambda(k))} \exp(\lambda(k)\boldsymbol{\mu}(k)^{\mathsf{T}}\boldsymbol{x}(n)), \tag{5}$$

where $\boldsymbol{\mu}(k)$ denotes the mean direction, $||\boldsymbol{\mu}(k)|| = 1$, and $\lambda(k)$ the concentration parameter, $\lambda(k) \geq 0$. The normalisation constant is calculated as

$$Z(\lambda(k)) = \frac{(2\pi)^{D/2} I_{D/2-1}(\lambda(k))}{\lambda(k)^{D/2-1}} \tag{6}$$

where $I_\nu(u)$ denotes the modified Bessel function of the first kind and order ν [5, 10.25]. The distribution parameters are modelled as random variables with prior distribution $p(\boldsymbol{\phi}(k)) = p(\boldsymbol{\mu}(k)|\lambda(k))p(\lambda(k))$ parametrised as proposed in [2]: $p(\boldsymbol{\mu}(k))$ is a VMF distribution with mean direction $\boldsymbol{m}_0$ and concentration parameter $\beta_0\lambda(k)$ while $p(\lambda(k))$ is a gamma distribution with shape parameter a_0 and inverse scale parameter b_0.

3 Variational inference

The previous section introduced the infinite mixture that we would like to associate with observations $\{\boldsymbol{x}(n)\}$. The current section focusses on the unobserved model parameters $\boldsymbol{w}$ and posterior distribution $p(\boldsymbol{w}|\{\boldsymbol{x}(n)\})$. The posterior distribution cannot be computed in closed-form, but we can approximate the conditional distribution $p(\boldsymbol{w}|\{\boldsymbol{x}(n)\})$ with a variational distribution $q(\boldsymbol{w})$ [6]. Variational methods convert posterior estimation into an optimisation problem: the variational distribution parameters are chosen to maximise the evidence lower bound (ELBO) $\mathbb{E}[\ln p(\boldsymbol{w}, \{\boldsymbol{x}(n)\})] - \mathbb{E}[\ln q(\boldsymbol{w})]$, where the expectations are calculated based on the variational distribution which is factorised to make the calculations tractable.

Factorisation means that the unobserved variables are partitioned and each subset $\boldsymbol{w}(j)$ (one or more unobserved variables) is associated with a variational distribution $q_j(\boldsymbol{w}(j))$ so that the complete variational posterior distribution is determined as $q(\boldsymbol{w}) = \prod_j q(\boldsymbol{w}(j))$, where $q(\boldsymbol{w}(j)) = q_j(\boldsymbol{w}(j))$. The variational distribution that maximises ELBO with respect to $\boldsymbol{w}(j)$ is then determined as

$$\ln q^*(\boldsymbol{w}(j)) = \mathbb{E}[\ln p(\boldsymbol{w}, \{\boldsymbol{x}(n)\})]_{i \neq j} + constant, \tag{7}$$

where $\mathbb{E}[f(\boldsymbol{w})]_{i \neq j}$ means that the expectation is calculated with respect to the unobserved parameters other than the current parameter $\boldsymbol{w}(j)$. Thus it is common that the optimal distribution $\ln q^*(\boldsymbol{w}(j))$ depends on the variational distributions associated with other parameters and iterative updates are needed to optimise the variational distribution parameters.

3.1 Joint distribution

Since the observations $\boldsymbol{x}(n)$ are independent when conditioned on $z(n)$ and distribution parameters $\boldsymbol{\phi}$ and the allocations $z(n)$ are independent when conditioned on the stick proportions $\boldsymbol{v}$, log-probabilities over the complete observation set and model parameters can be written as

$$\ln p(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\phi}, \{\boldsymbol{x}(n)\}) = \sum_n \ln p(\boldsymbol{x}(n)|z(n), \boldsymbol{\phi}) + \sum_n \ln p(z(n)|\boldsymbol{v}) + \ln p(\boldsymbol{v}) + \ln p(\boldsymbol{\phi}). \tag{8}$$

Equation (8) indicates that the joint probabilities can be calculated based on the observation likelihoods in Equation (1), allocation probabilities in Equation (4), and prior probabilities associated with stick proportions and observation model parameters.

The observations $\boldsymbol{x}(n)$ and unobserved allocations $z(n)$ are local variables associated with the individual observations n. The observation likelihoods are calculated as (Equation 1)

$$\ln p(\boldsymbol{x}(n)|z(n), \boldsymbol{\phi}) = \sum_{k=1}^{\infty} \mathbf{1}[z(n) = k] \ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k)), \tag{9}$$

where $p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))$ are the component-conditioned observation likelihoods determined in Equation (5). The component-conditioned likelihoods are calculated as

$$\ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k)) = \lambda(k)\boldsymbol{\mu}(k)^{\mathsf{T}}\boldsymbol{x}(n) - (\nu+1)\ln(2\pi) - \ln I_\nu(\lambda(k)) + \nu \ln \lambda(k), \tag{10}$$

where $\boldsymbol{\mu}(k)$ denotes the mean direction and $\lambda(k)$ the concentration parameter in VMF distribution and $\nu = D/2 - 1$. The allocation probabilities are calculated as (Equation 4)

$$\ln p(z(n)|\boldsymbol{v}) = \sum_{k=1}^{\infty} \mathbf{1}[z(n) > k] \ln(1 - v(k)) + \sum_{k=1}^{\infty} \mathbf{1}[z(n) = k] \ln v(k) \tag{11}$$

where $v(k)$ are stick proportions. The stick proportions and observation model parameters are associated with individual mixture components k rather than

observations n. The stick proportions $v(k)$ follow a beta distribution $Beta(1, \alpha)$ so that their complete distribution

$$\ln p(\boldsymbol{v}) = \sum_{k=1}^{\infty}[(\alpha - 1)\ln(1 - v(k)) + \ln\Gamma(1 + \alpha) - \ln\Gamma(\alpha)], \tag{12}$$

where $\Gamma(t)$ denotes the gamma function. The observation model parameters include parameters to control the mean direction $\boldsymbol{\mu}(k)$ and concentration $\lambda(k)$ so that

$$\ln p(\boldsymbol{\phi}) = \sum_{k=1}^{\infty}[\ln p(\boldsymbol{\mu}(k)|\lambda(k)) + \ln p(\lambda(k))], \tag{13}$$

where $p(\boldsymbol{\mu}(k)|\lambda(k))$ is modelled as VMF distribution and $p(\lambda(k))$ as gamma distribution:

$$\ln p(\boldsymbol{\mu}(k)|\lambda(k)) = \beta_0\lambda(k)\boldsymbol{m}_0^{\mathsf{T}}\boldsymbol{\mu}(k) - (\nu + 1)\ln(2\pi) - \ln I_\nu(\beta_0\lambda(k)) + \nu\ln\beta_0\lambda(k), \tag{14}$$

$$\ln p(\lambda(k)) = (a_0 - 1)\ln\lambda(k) - b_0\lambda(k) + a_0\ln b_0 - \ln\Gamma(a_0), \tag{15}$$

The variational posterior distribution discussed in the next section is factorised so that the expectation over the log-likelihood in Equation (8) can be computed based on expectations $\mathbb{E}[\ln p(\boldsymbol{x}(n)|z(n), \boldsymbol{\phi})]$, $\mathbb{E}[\ln p(z(n)|v)]$, $\mathbb{E}[\ln p(\boldsymbol{v})]$, and $\mathbb{E}[\ln p(\boldsymbol{\phi})]$ that are computed based on variational distributions associated with the unobserved variables. Expectations calculated based on $q(z(n))$ include $\mathbb{E}[\mathbf{1}(z(n) = k)]$ in Equation (9) and Equation (11), expectations calculated based on $q(v(k))$ include $\mathbb{E}[\ln v(k)]$ and $\mathbb{E}[\ln(1 - v(k))]$ in Equation (11) and Equation (12), and expectations calculated based on $q(\boldsymbol{\phi}(k))$ include $\mathbb{E}[\lambda(k)]$, $\mathbb{E}[\ln\lambda(k)]$, $\mathbb{E}[\ln I_\nu(\lambda(k))]$ and $\mathbb{E}[\lambda(k)\boldsymbol{\mu}(k)]$ in Equation (10) and Equation (13).

3.2 Variational distribution

The expectations presented in the previous section become computable when we assume a variational posterior distribution that is factored and truncated as proposed in [1]. This means that the latent variables $v(k)$, $\boldsymbol{\phi}(k)$, and $z(n)$ are assumed independent and the stick-breaking representation is truncated at truncation limit T so that the complete variational distribution

$$q(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\phi}) = \prod_{n=1}^{N} q(z(n)) \prod_{k=1}^{T-1} q(v(k)) \prod_{k=1}^{T} q(\boldsymbol{\phi}(k)), \tag{16}$$

where each unobserved variable $z(n)$, $v(k)$, and $\boldsymbol{\phi}(k)$ is associated with a separate distribution. Distributions $q(z(n))$ and $q(v(k))$ are constructed as in [1] while the variational distribution $q(\boldsymbol{\phi}(k))$ is close to the variational distribution proposed in [2].

3.2.1 $q(z(n))$

To determine the variational distribution associated with the discrete indicator variable $z(n)$, we collect terms that depend on $z(n) = k$ in Equation (8). These include the nth term in the first and second summation,

$$\ln q^*(z(n)) = \mathbb{E}[\ln p(\boldsymbol{x}(n)|z(n), \boldsymbol{\phi})] + \mathbb{E}[\ln p(z(n)|\boldsymbol{v})] + cst, \tag{17}$$

where expectation values are calculated based on the variational distributions $q(v(k))$ and $q(\boldsymbol{\phi}(k))$ and cst includes terms that do not depend on $z(n)$. The expectation values are determined as

$$\mathbb{E}[\ln p(\boldsymbol{x}(n)|z(n),\boldsymbol{\phi})] = \sum_{k=1}^{T} \mathbf{1}[z(n)=k]\mathbb{E}[\ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))], \tag{18}$$

$$\mathbb{E}[\ln p(z(n)|\boldsymbol{v})] = \sum_{k=1}^{T} \mathbf{1}[z(n)=k]\mathbb{E}[\ln v(k)] + \sum_{k=1}^{T} \mathbf{1}[z(n)>k]\mathbb{E}[\ln(1-v(k))], \tag{19}$$

where the infinite summations in Equation (9) and Equation (11) are truncated because we assumed a truncated variational distribution that does not allocate observations to components $k > T$. The second summation in Equation (19) can be written as

$$\sum_{k=1}^{T} \mathbf{1}[z(n)>k]\mathbb{E}[\ln(1-v(k))] = \sum_{k=1}^{T} \left[\mathbf{1}[z(n)=k] \sum_{i=1}^{k-1} \mathbb{E}[\ln(1-v(i))] \right]. \tag{20}$$

Thus it is possible to combine the summations in Equation (18) and Equation (19) and reorganise Equation (17) around the common denominator $\mathbf{1}[z(n)=k]$:

$$\ln q^*(z(n)) = \sum_{k=1}^{T} \mathbf{1}[z(n)=k]\mathbb{E}[\ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k)) + \ln v(k) + \sum_{i=1}^{k-1} \ln(1-v(i))] + cst. \tag{21}$$

This is recognised as a categorical distribution with event probabilities $\gamma(n,k)$:

$$\ln q^*(z(n)) = \sum_{k=1}^{T} \mathbf{1}[z(n)=k] \ln \gamma(n,k), \tag{22}$$

where the probabilities $\gamma(n,k) \propto \exp(s(n,k))$ and

$$s(n,k) = \mathbb{E}[\ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))] + \mathbb{E}[\ln v(k)] + \sum_{i=1}^{k-1} \mathbb{E}[\ln(1-v(i))]. \tag{23}$$

The event probabilities $\gamma(n,k)$ are normalised so that probabilities associated with each observation n sum to 1. When expectations are calculated under the variational posterior distribution in Equation (22), the indicator function $\mathbf{1}[z(n)=k]$ has expectation value $\mathbb{E}[\mathbf{1}[z(n)=k]] = \gamma(n,k)$ and the variational distribution $\mathbb{E}[(q^*(z(n)))] = \sum_{k=1}^{T} \gamma(n,k) \ln \gamma(n,k)$.

3.2.2 $q(v(k))$

The variational distribution associated with stick proportion $v(k)$, $k < T$, is determined as proposed in [1]. We rewrite the right-hand side in Equation (8) as

$$\sum_{n} \mathbf{1}[z(n)=k] \ln v(k) + \sum_{n} \mathbf{1}[z(n)>k] \ln(1-v(k)) + (\alpha-1)\ln(1-v(k)) + cst \tag{24}$$

where cst includes terms that do not depend on $v(k)$ and α is the prior distribution parameter in Equation (12). To determine the variational distribution $q^*(v(k))$, we calculate the expectation based on $q^*(z(n))$ and combine the terms that include $\ln(1 - v(k))$:

$$\ln q^*(v(k)) = [\sum_n \gamma(n,k)] \ln v(k) + ([\sum_n \sum_{i>k} \gamma(n,i)] + \alpha - 1) \ln(1 - v(k)) + cst, \tag{25}$$

which is recognised as a beta distribution with parameters $g_1(k)$ and $g_2(k)$,

$$\ln q^*(v(k)) = (g_1(k)-1) \ln v(k) + (g_2(k)-1) \ln(1-v(k)) - \ln B(g_1(k), g_2(k) \tag{26}$$

where $\ln B(g_1, g_2) = \ln \Gamma(g_1) + \ln \Gamma(g_2) - \ln \Gamma(g_1 + g_2)$ and the distribution parameters are calculated as

$$g_1(k) = 1 + \sum_n \gamma(n,k), \tag{27}$$

$$g_2(k) = \alpha + \sum_n \sum_{i>k} \gamma(n,i). \tag{28}$$

The variational distribution is used to calculate $\mathbb{E}[\ln v(k)]$ and $\mathbb{E}[\ln(1 - v(k))]$ that are substituted in Equation (23): $\mathbb{E}[\ln v(k)] = \psi(g_1(k)) - \psi(g_1(k) + g_2(k))$, where $\psi(x)$ denotes the standard digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$, and $\mathbb{E}[\ln(1 - v(k))] = \psi(g_2(k)) - \psi(g_1(k) + g_2(k))$. This is because $1 - v(k)$ calculated based on a beta-distributed variable $v(k)$ with distribution $Beta(g_1(k), g_2(k))$ is also beta-distributed with distribution $Beta(g_2(k), g_1(k))$. To evaluate the evidence lower bound we also calculate

$$\mathbb{E}[q^*(v(k))] = (g_1 - 1)\psi(g_1) + (g_2 - 1)\psi(g_2) - (g_1 + g_2 - 2)\psi(g_1 + g_2) - \ln B(g_1, g_2), \tag{29}$$

where $g_1 = g_1(k)$ and $g_2 = g_2(k)$.

3.2.3 $q(\boldsymbol{\phi}(k))$

The variational distribution associated with component-conditioned observation model parameters $\boldsymbol{\phi}(k)$ is determined based on the terms in Equation (8) that depend on $\boldsymbol{\phi}(k)$,

$$\ln q^*(\boldsymbol{\phi}(k)) = \sum_n \gamma(n,k) \ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k)) + \ln p(\boldsymbol{\phi}(k)) + cst. \tag{30}$$

The variational distribution can be factorised as $q(\boldsymbol{\phi}(k)) = q(\boldsymbol{\mu}(k)|\lambda(k))q(\lambda(k))$, as proposed in [2]. The variational distribution $q^*(\boldsymbol{\mu}(k)|\lambda(k))$ is fixed based on the terms that depend on the mean direction $\boldsymbol{\mu}(k)$,

$$\ln q^*(\boldsymbol{\mu}(k)|\lambda(k)) = \sum_n [\gamma(n,k)\lambda(k)\boldsymbol{\mu}(k)^\mathsf{T}\boldsymbol{x}(n)] + \beta_0\lambda(k)\boldsymbol{m}_0^\mathsf{T}\boldsymbol{\mu}(k) + cst, \tag{31}$$

where $\boldsymbol{m}_0$ is the mean direction and β_0 the scale applied on the concentration parameter in the prior distribution in Equation (14). Equation (31) can be reorganised into

$$\ln q^*(\boldsymbol{\mu}(k)|\lambda(k)) = \lambda(k)(\beta_0\boldsymbol{m}_0 + \sum_n \gamma(n,k)\boldsymbol{x}(n))^\mathsf{T}\boldsymbol{\mu}(k) + cst, \tag{32}$$

which is recognised as a VMF distribution with mean direction $\boldsymbol{m}(k)$ and concentration parameter $\beta(k)\lambda(k)$,

$$\ln q^*(\boldsymbol{\mu}(k)|\lambda(k)) = \beta(k)\lambda(k)\boldsymbol{m}(k)^\mathsf{T}\boldsymbol{\mu}(k) - \ln Z(\lambda(k)), \tag{33}$$

where $\ln Z(k) = (\nu+1)\ln(2\pi)+\ln I_\nu(\beta(k)\lambda(k))-\nu\ln(\beta(k)\lambda(k))$ with $\nu = D/2-1$ and the mean direction and scale parameter calculated as $\boldsymbol{m}(k) = \boldsymbol{r}(k)/||\boldsymbol{r}(k)||$ and $\beta(k) = ||\boldsymbol{r}(k)||$ where

$$\boldsymbol{r}(k) = \beta_0\boldsymbol{m}_0 + \sum_n \gamma(n,k)\boldsymbol{x}(n). \tag{34}$$

The variational distribution associated with concentration parameter $\lambda(k)$ is then determined based on

$$\ln q^*(\lambda(k)) = \ln q^*(\boldsymbol{\phi}(k)) - \ln q^*(\boldsymbol{\mu}(k)|\lambda(k)), \tag{35}$$

where $\ln q^*(\boldsymbol{\phi}(k))$ is the distribution in Equation (30) and $q^*(\boldsymbol{\mu}(k)|\lambda(k))$ the distribution in Equation (33). The distribution is fixed based on the terms that depend on $\lambda(k)$:

$$(\nu N(k)+a_0-1)\ln\lambda(k)-b_0\lambda(k)-N(k)\ln I_\nu(\lambda(k))-\ln I_\nu(\beta_0\lambda(k))+\ln I_\nu(\beta(k)\lambda(k)), \tag{36}$$

where a_0 and b_0 are the prior distribution parameters in Equation (15) and we introduce $N(k) = \sum_n \gamma(n,k)$. Here we approximate $\ln I_\nu(\lambda(k))$, $\ln I_\nu(\beta_0\lambda(k))$ and $\ln I_\nu(\beta(k)\lambda(k))$ with upper and lower bounds as proposed in [2]. Approximation is based on the observation that $\ln I_\nu(x)$ ($\nu > 0$) is concave with respect to x and convex relative to $\ln x$ ($x > 0$) [2, Lemma 2.1–2.2]. A concave and differentiable function $f(x)$ has an upper bound

$$f(x) \leq f(a) + f'(a)(x-a) \tag{37}$$

where $f'(a)$ denotes a differential calculated with respect to x and evaluated at linearisation point a. The second observation provides a lower bound since a differentiable function $f(x)$ that is convex relative to function $g(x)$ can be bounded as

$$f(x) \geq f(a) + \frac{f'(a)}{g'(a)}(g(x)-g(a)), \tag{38}$$

The approximations are used to construct a lower bound to the evidence lower bound: $\ln I_\nu(\lambda(k))$ and $\ln I_\nu(\beta_0\lambda(k))$ that decrease ELBO are substituted with the upper bound in Equation (37) while $\ln I_\nu(\beta(k)\lambda(k))$ that increases ELBO is substituted with the lower bound in Equation (38) [2, Equation (19)–(21)]. The differential is calculated as

$$f'(x) = \frac{I'_\nu(x)}{I_\nu(x)}, \tag{39}$$

where $I'_\nu(x)$ is calculated with the recurrence relation $I'_\nu(x) = I_{\nu+1}(x) + \frac{\nu}{x}I_\nu(x)$ [5, 10.29.2]. To compute the lower bound, we also determine $g'(x) = 1/x$ when $g(x) = \ln x$. The upper and lower bounds are substituted in Equation (36), terms that do not depend on $\lambda(k)$ are discarded, and terms that depend on

$\ln \lambda(k)$ or $\lambda(k)$ are reorganised so that expression based on which the variational distribution $q^*(\lambda(k))$ is determined becomes

$$(\nu N(k) + \beta(k)\bar{\lambda} f'(\beta(k)\bar{\lambda}) + a_0 - 1)\ln \lambda(k) - (b_0 + N(k) f'(\bar{\lambda}) + \beta_0 f'(\beta_0 \bar{\lambda}))\lambda(k), \tag{40}$$

where $\bar{\lambda} = \bar{\lambda}(k)$ is the linearisation point and $f'(x)$ is calculated as in Equation (39). This corresponds to a gamma distribution with shape parameter $a(k)$ and inverse scale parameter $b(k)$

$$\ln q^*(\lambda(k)) = (a(k) - 1)\ln \lambda(k) - b(k)\lambda(k) + a(k)\ln b(k) - \ln \Gamma(a(k)), \tag{41}$$

with the parameters calculated as

$$a(k) = a_0 + \nu N(k) + \beta(k)\bar{\lambda}(k) f'(\beta(k)\bar{\lambda}(k)), \tag{42}$$

$$b(k) = b_0 + N(k) f'(\bar{\lambda}(k)) + \beta_0 f'(\beta_0 \bar{\lambda}(k)). \tag{43}$$

The variational posterior distribution is used to calculate expectation over Equation (10) that is substituted in Equation (23). The gamma distribution in Equation (41) provides closed-form solutions for $\mathbb{E}[\lambda(k)] = a(k)/b(k)$ and $\mathbb{E}[\ln \lambda(k)] = \psi(a(k)) - \ln b(k)$, while $\mathbb{E}[\ln I_\nu(\lambda(k))]$ is approximated with expectation calculated over the upper bound which is linear with respect to $\lambda(k)$ (Equation 37). Expectations that need to be calculated based on the posterior also include $\mathbb{E}[\lambda(k)\boldsymbol{\mu}(k)]$. This does not have a closed-form solution due to the relation between $\boldsymbol{\mu}(k)$ and $\lambda(k)$: $\mathbb{E}[\boldsymbol{\mu}(k)] = \mathbb{E}[I_{\nu+1}(\beta(k)\lambda(k))/I_\nu(\beta(k)\lambda(k))]\boldsymbol{m}(k)$. The expectation corresponds to mean direction $\boldsymbol{m}(k)$ multiplied with an expression that is zero when the distribution is unconcentrated ($\beta(k)\lambda(k) = 0$) and approaches one when the concentration parameter $\beta(k)\lambda(k) \to \infty$. We assume that the variational posterior distribution $q^*(\boldsymbol{\mu}(k)|\lambda(k))$ is concentrated around $\boldsymbol{m}(k)$ and calculate the variational distributions $q^*(z(n))$ based on approximate likelihoods

$$\mathbb{E}[\ln p(\boldsymbol{x}(n)|\boldsymbol{\phi}(k))] \approx \mathbb{E}[\lambda(k)]\boldsymbol{m}(k)^\mathsf{T}\boldsymbol{x}(n) + (\nu + 1)\ln(2\pi) - \bar{f}(\lambda(k)) + \nu\mathbb{E}[\ln \lambda(k)], \tag{44}$$

where $\bar{f}(\lambda(k))$ denotes expectation calculated over the linearisation that upper-bounds $\ln I_\nu(\lambda(k))$, $\bar{f}(\lambda(k)) = \ln I_\nu(\bar{\lambda}(k)) + f'(\bar{\lambda}(k))(\mathbb{E}[\lambda(k)] - \bar{\lambda}(k))$.

The variational posterior distribution discussed in this section maximises a lower bound to the evidence lower bound as proposed in [2]. To evaluate the lower bound, we must determine $\mathbb{E}[\ln q^*(\boldsymbol{\phi}(k))]$ which is calculated based on $\mathbb{E}[\ln q^*(\boldsymbol{\mu}(k)|\lambda(k))]$ and $\mathbb{E}[\ln q^*(\lambda(k))]$. Expectation over the variational distribution associated with the concentration parameter has a closed-form solution,

$$\mathbb{E}[\ln q^*(\lambda(k))] = (a(k) - 1)\psi(a(k)) + a(k) - \ln b(k) + \ln \Gamma(a(k)), \tag{45}$$

whereas expectation over the variational distribution associated with mean direction:

$$\mathbb{E}[\ln q^*(\boldsymbol{\mu}(k)|\lambda(k))] = \beta(k)\mathbb{E}[\lambda(k)\boldsymbol{\mu}(k)]^\mathsf{T}\boldsymbol{m}(k) - \mathbb{E}[\ln Z(\beta(k)\lambda(k))], \tag{46}$$

where $\mathbb{E}[\ln Z(\beta(k)\lambda(k))]$ includes the intractable moment $\mathbb{E}[\ln I_\nu(\beta(k)\lambda(k))]$. The expectations are computed based a simple approximation: $I_\nu(\beta(k)\lambda(k))$ is substituted with a scaled exponential $\exp(\beta(k)\lambda(k))/\sqrt{2\pi\beta(k)\lambda(k)}$ [5, 10.30]. When the exponential approximation is used, $\mathbb{E}[\lambda(k)\boldsymbol{\mu}(k)] = \mathbb{E}[\lambda(k)]\boldsymbol{m}(k)$ and $\mathbb{E}[\ln I_\nu(\beta(k)\lambda(k))] = \beta(k)\mathbb{E}[\lambda(k)] - 0.5\mathbb{E}[\ln \lambda(k)] - 0.5\ln \beta(k) - 0.5\ln(2\pi)$. The exponential approximation is also used to calculate $\mathbb{E}[\ln p(\boldsymbol{\mu}(k)|\lambda(k))]$ (Equation 14).

3.3 Numerical issues

The approximations that were utilised to derive the variational distribution do not prevent numerical issues related to $I_\nu(x)$ which is evaluated with the MATLAB function `besseli`. The function is evaluated when we calculate the posterior distribution parameters in Equation (41) and expected likelihoods in Equation (44). To be precise, the posterior distribution parameters and expected likelihoods depend on the differential in Equation (39) and this is evaluated at $\bar{\lambda}(k)$, $\beta_0\bar{\lambda}(k)$, and $\beta(k)\bar{\lambda}(k)$. The differential is calculated as $f'(x) = r_\nu(x)+\nu/x$, where $r_\nu(x) = I_{\nu+1}(x)/I_\nu(x)$. When `besseli` does not produce a finite value, we approximate $r_\nu(x)$ with the simple bounds proposed in [7, Equation (9)]. The bounds which are applicable to $r_\nu(x)$ solve the numerical issues related to posterior parameter calculation, but the expected likelihoods in Equation (44) depend on the complete upper bound $\bar{f}(\lambda(k))$ that also includes $I_\nu(x)$ evaluated at $\bar{\lambda}(k)$. Since we are free to choose the linearisation point $\bar{\lambda}(k)$, we can avoid numerical issues by constraining $\bar{\lambda}(k)$ not to exceed the limit where `besseli` is finite. $\beta_0\bar{\lambda}(k)$ and $\beta(k)\bar{\lambda}(k)$ can exceed $\bar{\lambda}(k)$, but $I_\nu(x)$ does not need to be evaluated at $\beta_0\bar{\lambda}(k)$ or $\beta(k)\bar{\lambda}(k)$ since we calculate $\mathbb{E}[\ln p(\boldsymbol{\phi}(k))]$ and $\mathbb{E}[\ln q(\boldsymbol{\phi}(k))]$ based on the exponential approximation [5, 10.30].

References

[1] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[2] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, 2014.

[3] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, p. 161, 2001.

[4] K. V. Mardia and P. E. Jupp, *Directional Statistics*, vol. 494. John Wiley & Sons, 2009.

[5] "NIST digital library of mathematical functions." `http://dlmf.nist.gov/`, Release 1.0.13 of 2016-09-16. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[6] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," 2016, `arxiv.org/abs/1601.00670`.

[7] D. E. Amos, "Computation of modified Bessel functions and their ratios," *Math. Comp.*, vol. 28, no. 125, pp. 239–251, 1974.