Bayesian Inference with the von-Mises-Fisher Distribution in 3D

Julian Straub
jstraub@csail.mit.edu
http://people.csail.mit.edu/jstraub/

Abstract

In this writeup, I give an introduction to the von-Mises-Fisher (vMF) distribution which is a commonly used isotropic distribution for directional data. The writeup is an excerpt of my PhD thesis [10] with a focus on Bayesian inference and computational considerations when working with the vMF distribution. While the initial discussion is general, some of the results and derivations for efficient inference are specialized to 3D directional data. Specifically, after the introduction of the vMF distribution and two different conjugate prior distributions, I outline general sampling from the posterior vMF distribution before deriving the normalization of the prior and the marginal data distribution for 3D. The last two sections show the cumulative density function and the entropy for the 3D vMF distribution.

The von-Mises-Fisher (vMF) distribution is commonly used to describe directional data [2, 3, 6, 11] and can be regarded as akin to the isotropic Gaussian distribution of the sphere in D dimensions, $\mathbb{S}^{D-1}$. It is parametrized by a mean direction $\mu \in \mathbb{S}^{D-1}$ and a concentration $\tau > 0$ (see Fig. 1). Its density is defined as [4]

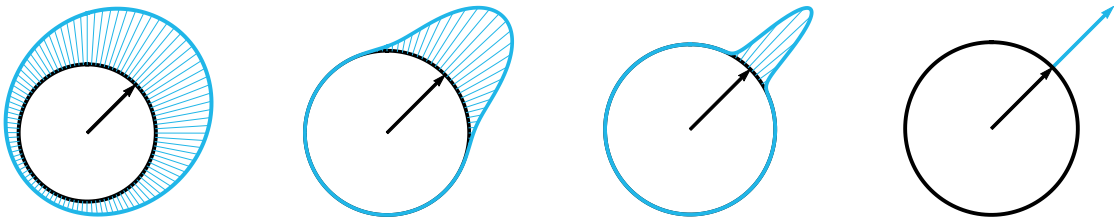
$$\text{vMF}(n;\mu,\tau) = Z(\tau)\exp(\tau\mu^T n)\,, \quad Z(\tau) = (2\pi)^{-D/2}\frac{\tau^{D/2-1}}{\text{I}_{D/2-1}(\tau)}\,, \tag{1}$$

where I_ν is the modified Bessel function [1] of the first kind of order ν. Figure 1 and 2 illustrates the vMF distribution in 2D and 3D respectively for different concentration parameters. In $D = 3$ dimensions, with $\frac{\tau^{1/2}}{\text{I}_{1/2}(\tau)} = \sqrt{\frac{\pi}{2}}\frac{\tau}{\sinh(\tau)}$ and $\sinh\tau = \frac{\exp\tau - \exp(-\tau)}{2}$, the normalizer of the vMF distribution simplifies to

$$Z(\tau) = \frac{\tau}{4\pi\sinh(\tau)} = \frac{\tau}{2\pi(\exp\tau - \exp(-\tau))}\,. \tag{2}$$

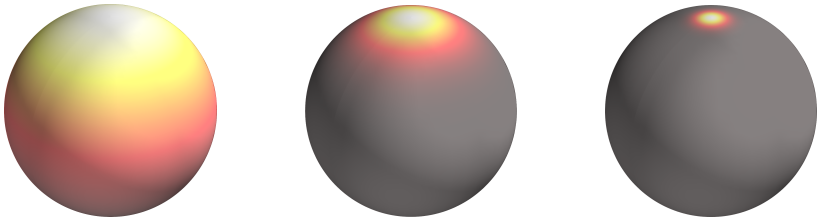
A numerically more stable way of writing the vMF distribution in 3D is

$$\text{vMF}(n;\mu,\tau) = \frac{\tau\exp\left(\tau\left(\mu^T n - 1\right)\right)}{2\pi\left(1-\exp(-2\tau)\right)}\,. \tag{3}$$



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 100$ (d) $\tau \to \infty$

Figure 1: Depiction of 2D von-Mises-Fisher distributions with increasing concentration τ. As $\tau \to \infty$ the von-Mises-Fisher distribution approaches a delta function on the sphere at its mode μ.



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 100$

Figure 2: The von-Mises-Fisher distributions on the unit sphere in 3D, $\mathbb{S}^2$, with mean at the north pole and concentrations τ. The color encodes the probability density function value of the vMF over the whole sphere. From the coloring it can be observed that the von-Mises-Fisher distribution is isotropic.

1. Sampling

An efficient approach for sampling from a vMF distribution was describe by Ulrich [12]. While the approach works for any dimension we describe it for $D = 3$ dimensions for clarity reasons. To sample a random vector from a vMF distribution with mode $m = (0, 0, 1)$ first sample the two variables u and v:

$$v \sim \text{Unif}(\mathbb{S}^{D-2}) \tag{4}$$

$$u \sim p(u; \tau) = \frac{\tau}{2 \sinh \tau} \exp(\tau u) \tag{5}$$

and then compute the vMF distributed sample n as

$$n = \begin{pmatrix} \sqrt{1 - u^2} v & u \end{pmatrix} . \tag{6}$$

In practice we obtain v by sampling from a zero-mean isotropic Gaussian with unit variance and normalizing the resulting sample to unit length. The inversion method is used to sample u (see general treatment in Sec. 2.1.2 of [9]): With the cumulative density of u, $F(u)$, derived from $p(u; \tau)$

$$\xi \sim \text{Unif}(0, 1) \tag{7}$$

$$u = F^{-1}(\xi) = 1 + \tau^{-1} \log \left(\xi + (1 - \xi) \exp(-2\tau) \right) . \tag{8}$$

Finally, we rotate the sampled vector n from m to μ via the rotation ${}^{\mu}R_m$ which can be computed from axis $w = m \times \mu$ and angle $\theta = \arccos(\mu^T m)$. With $\omega = \theta w$:

$${}^{\mu}R_m = \text{Exp}([\omega]_\times) = \text{I} + \tfrac{\sin(\theta)}{\theta} [\omega]_\times + \tfrac{1 - \cos(\theta)}{\theta^2} [\omega]_\times{}^2 \tag{9}$$

$$[\omega]_\times = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} . \tag{10}$$

This makes the overall sampling of a vMF distributed data point very efficient for $D = 3$ dimensions and the computational complexity independent of the concentration τ (which rejection sampling is not for example).

2. Conjugate Prior of μ given τ

For Bayesian inference it is convenient to have conjugate priors for the parameters of a distribution because posterior distributions remain in the same class as the prior distribution. For the vMF distribution mean parameter μ the conjugate prior, given a fixed τ, is a vMF distribution $\text{vMF}(\mu; \mu_0, \tau_0)$ [8]. Note that setting τ_0 to 0 amounts to assuming an uniform prior distribution for the mean μ. The corresponding posterior given directional data $\mathbf{n} = \{n_i\}_{i=1}^N$ is

$$\begin{aligned} p(\mu \mid \mathbf{n}; \tau, \mu_0, \tau_0) &\propto \text{vMF}(\mu; \mu_0, \tau_0) \prod_{i=1}^{N} \text{vMF}(n_i; \mu, \tau) \\ &= Z(\tau_0) Z(\tau)^N \exp\left(\mu^T \left(\tau_0 \mu_0 + \tau \sum_{i=1}^{N} n_i \right) \right) . \end{aligned} \tag{11}$$

The last expression has the form of a vMF distribution in μ and thus:

$$p(\mu \mid \mathbf{n}; \tau, \mu_0, \tau_0) = \text{vMF}\left(\mu; \tfrac{\vartheta_N}{\|\vartheta_N\|_2}, \|\vartheta_N\|_2\right) , \tag{12}$$

where $\vartheta_N = \tau_0\mu_0 + \tau \sum_{i=1}^N n_i$. Under the conjugate prior for μ we can compute the marginalization in closed form as:

$$\begin{aligned} p(n; \tau, \mu_0, \tau_0) &= \int_{\mathbb{S}^{D-1}} \text{vMF}(n; \mu, \tau)\, \text{vMF}(\mu; \mu_0, \tau_0)\, \mathrm{d}\mu \\ &= Z(\tau)Z(\tau_0) \int_{\mathbb{S}^{D-1}} \exp\left(\mu^T(\tau n + \tau_0\mu_0)\right) \mathrm{d}\mu \\ &= Z(\tau)Z(\tau_0) \int_{\mathbb{S}^{D-1}} \exp\left(\|\vartheta_1\|_2 \mu^T \tfrac{\vartheta_1}{\|\vartheta_1\|_2}\right) \mathrm{d}\mu \\ &= \frac{Z(\tau)Z(\tau_0)}{Z(\|\vartheta_1\|_2)} = \frac{Z(\tau)Z(\tau_0)}{Z(\|\tau n + \tau_0\mu_0\|_2)} , \end{aligned} \tag{13}$$

where $\vartheta_1 = \tau n + \tau_0\mu_0$ as introduced in Eq. (12).

3. Joint Conjugate Prior for μ and τ

There also exists a conjugate prior distribution for the mean μ and the concentration parameter τ which unfortunately is only known up to proportionality [8]:

$$p(\mu, \tau \mid \mu_0, a, b) \propto \left(\frac{\tau^{D/2-1}}{\mathrm{I}_{D/2-1}(\tau)}\right)^a \exp\left(b\tau\mu^T\mu_0\right) , \tag{14}$$

where $0 < b < a$. The normalizing constant can only be computed analytically in special cases. Knowing this prior only up to proportionality still allows sampling from it for sampling-based inference. The posterior given observed data $\{n_i\}_{i=1}^N$ is

$$p(\mu, \tau \mid \{n_i\}_{i=1}^N, \mu_0, a, b) \propto \prod_{i=1}^N Z(\tau) \exp(n_i^T\mu\tau) \left(\frac{\tau^{D/2-1}}{\mathrm{I}_{D/2-1}(\tau)}\right)^a \exp\left(b\tau\mu^T\mu_0\right) \tag{15}$$

$$\propto \left(\frac{\tau^{D/2-1}}{\mathrm{I}_{D/2-1}(\tau)}\right)^{a+N} \exp\left(\tau\mu^T\left(\sum_{i=1}^N n_i + b\mu_0\right)\right) \tag{16}$$

$$= \left(\frac{\tau^{D/2-1}}{\mathrm{I}_{D/2-1}(\tau)}\right)^{a_N} \exp\left(\tau b_N \mu^T\mu_N\right) , \tag{17}$$

where the posterior parameters are

$$a_N = a + N\,, \qquad b_N = \|\vartheta\|_2\,, \qquad \mu_N = \lceil\vartheta\rceil\ , \qquad \vartheta = \sum_{i=1}^N n_i + b\mu_0\,. \tag{18}$$

Observe that $0 < b_N < a_N$ because of $0 < b < a$. This shows that a acts similar to the pseudo counts ν and κ of for example the normal inverse Wishart distribution [5].

3.1. Sampling from the Joint Prior

One way to sample from this prior distribution is using Gibbs sampling:

$$\mu \sim p(\mu|\tau; \mu_0, a, b) \propto \text{vMF}(\mu; \mu_0, b\tau) \tag{19}$$

$$\tau \sim p(\tau|\mu; \mu_0, a, b)\,. \tag{20}$$

Sampling μ amounts to sampling from a vMF distribution as described earlier in this section, while sampling from the conditional distribution of τ needs special care. Inversion sampling is not applicable since the cumulative density could not be inverted. Instead, we choose to use a slice sampler [7], an efficient sampling strategy for low-dimensional distributions. Since

Algorithm 1 Slice sampler for the prior distribution on the von-Mises-Fisher concentration τ.

Require: $f(\tau) \propto p(\tau|\mu; \mu_0, a, b)$ and τ_0
Find maximum $\tau^\star$ of $f(\tau)$
Initialize $\tau = \tau_0$
while more samples desired **do**
  Sample $u \sim \text{Unif}(0, f(\tau))$
  if $\tau^\star > 0$ **then**
    Find left slice border τ_L of $f(\tau)$ using Newton starting from $10^{-3}\tau^\star$
    Find right slice border τ_R of $f(\tau)$ using Newton starting from $1.5\tau^\star$
  else
    $\tau_L = 0$
    Find right slice border τ_R of $f(\tau)$ using Newton starting from 0.5
  end if
  Sample $\tau \sim \text{Unif}(\tau_L, \tau_R)$
end while

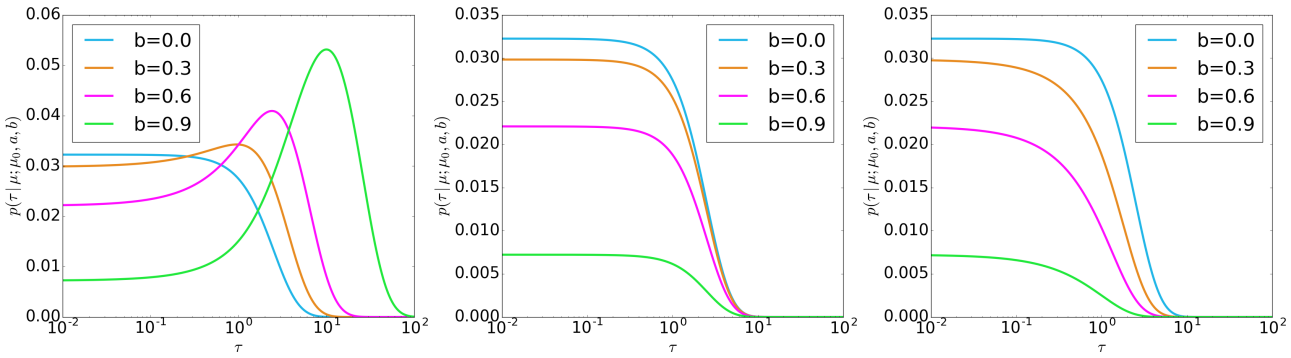


Figure 3: The conjugate prior for the parameters of a von-Mises-Fisher distribution is depicted for different values of b and dot-product $\mu^T\mu_0$. From left to right dot products of 1 (μ and μ_0 are equal), 0 (μ and μ_0 are orthogonal), and -1 (μ and μ_0 pointing in opposite directions) are shown. Hyper-parameters b are sampled in the allowed interval from 0 to 1.

the distribution is unimodal (see Fig. 3 and Appendix A.1 for a full characterization), a slice sampler can be implemented more efficiently than standard more universal algorithms explored in [7]. In the following we use $f(\tau) \propto p(\tau|\mu; \mu_0, a, b)$.

The slice sampler outlined in Alg. 1 alternates between sampling u uniformly from 0 to $f(\tau)$ and sampling τ uniformly from the set $\mathcal{T} = \{\tau : f(\tau) \geq u\}$. As discussed in depth in Appendix A.1, there are two cases for the set $\mathcal{T}$: either the maximum is attained at $\tau^\star = 0$ and the function decreases for $\tau > 0$ or the maximum is attained for some $\tau^\star > 0$ and the function increases for $\tau < \tau^\star$ and decreases for $\tau > \tau^\star$. In the first case $\mathcal{T}$ is the set from 0 to $f(\tau) = u$, which can be found efficiently using Newton's method starting from some arbitrary small $\tau_R^0 = 0.5$. In the second case $\mathcal{T}$ is set from τ_L to τ_R, where τ_L and τ_R are the locations of $f(\tau) = u$ left and right of the maximum. The two intersection points can be found by running Newton's algorithm from sufficiently far left/right of the maximum $\tau^\star$. In practice we start Newton's method from $\tau_L^0 = 10^{-3}\tau^\star$ to obtain the left intersection point and from $\tau_R^0 = 1.5\tau^\star$ to obtain the right intersection point. Starting closer to $\tau^\star$ leads to numerical problems. Newton's method generally converges in less than 10 iterations.

3.2. Normalization of the Joint Prior for $a = 1$ and $D = 3$

While we can directly sample from the joint prior using the aforementioned method, we do need a parametric normalized form for the evaluation of the marginal data distribution $p(x_i; \mu_0, a, b)$ for e.g. Bayesian nonparametric inference. The problem with finding a normalizer for the prior is largely due to the exponentiation with a. Setting $a = 1$ and working in $D = 3$ dimensions, we can derive a closed form normalizer as shown in Appendix A.2. Setting $a = 1$ amounts to assuming a weak prior since a can be thought of as pseudo counts as discussed in relation to the posterior parameter updates in Eq. (18). Using a weak prior is a common practice if there are is no strong prior information about the distribution of the parameters.

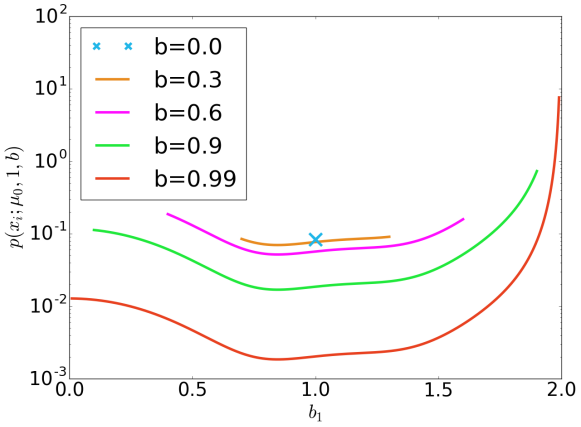


Figure 4: The marginal distribution of a von-Mises-Fisher distributed data point under the conjugate prior for $b_1 = \|x_i + b\mu_0\|_2$. Since a given value b restricts the range of $b_1 = \|x_i + b\mu_0\|_2$ the plots all have a different support.

With this the properly normalized prior for μ and τ is

$$p(\mu, \tau \mid \mu_0, 1, b) = \frac{b\tau}{2\pi^2} \frac{\exp\left(b\tau\mu^T\mu_0\right)}{\tan\left(\frac{b\pi}{2}\right)\sinh(\tau)} \,. \tag{21}$$

Slices of the prior density are plotted for aligned μ and μ_0 ($\mu^T\mu_0 = 1$), orthogonal μ and μ_0 ($\mu^T\mu_0 = 0$), and flipped μ and μ_0 ($\mu^T\mu_0 = -1$) and different b between 0 and 1 in Fig. 3. They show that the prior encourages a low concentration for unaligned μ and μ_0. Only once μ and μ_0 become aligned, the prior encourages higher concentrations. The magnitude of the most likely concentration then increases with b as can be seen from the left-most plot in Fig. 3.

3.3. Marginal Data Distribution

For $D = 3$ dimensions and $a = 1$ we can derive a closed form normalized probability density function for the marginal distribution of the data under the prior:

$$p(x_i; \mu_0, 1, b) = \int_0^\infty \int_{\mathbb{S}^2} \text{vMF}(x_i; \mu, \tau) p(\mu, \tau; \mu_0, 1, b) \,\mathrm{d}\mu \,\mathrm{d}\tau \tag{22}$$

$$= \frac{b}{2^3 \tan\left(\frac{b\pi}{2}\right)} \frac{1 - \text{sinc}(b_1\pi)}{\sin^2(\frac{b_1\pi}{2})} \,, \tag{23}$$

where $0 < b_1 = \|x_i + b\mu_0\|_2 < 2$. For the full derivation refer to Appendix A.3. This marginal distribution is displayed as a function of b_1 for several b values in Fig. 4. Since a given value b restricts the range of $b_1 = \|x_i + b\mu_0\|_2$ the plots all have a different support. For $b = 0$ the prior is uniform over the sphere, $b_1 = 1$ for all x_i and we therefore expect $p(x_i; \mu_0, 1, 0)$ to be equal to one over the area of the sphere $\mathbb{S}^2$ which is indeed the case.

4. Cumulative Density Function in 3D

The cumulative density function (cdf) of the radially symmetric vMF distribution is the probability that the angle between the mode μ and a data point n is smaller than α. Working in spherical coordinates and arbitrarily fixing $\mu = (0, 0, 1)$ the cdf

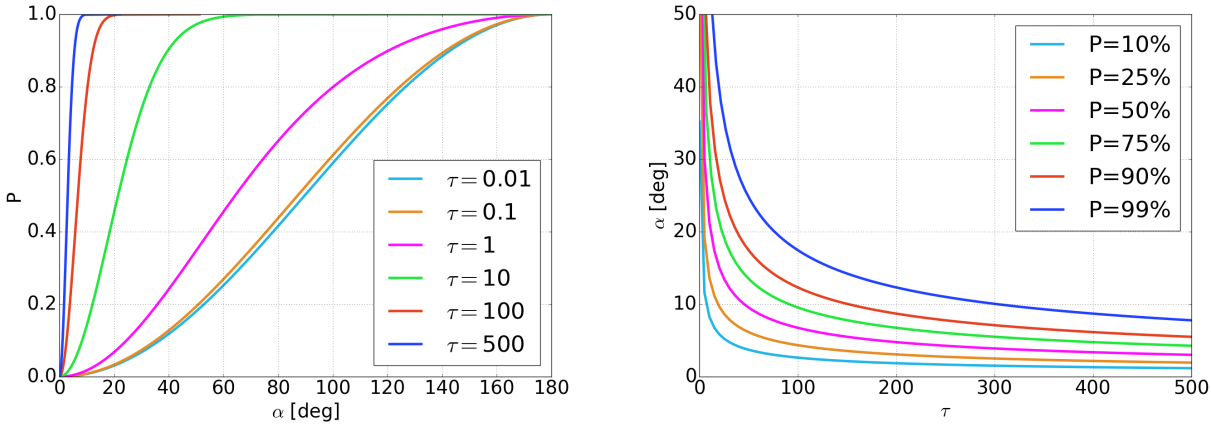


Figure 5: The left plot shows the cumulative density P of the vMF distribution as a function of the concentration τ depending on desired percentile P. On the right, the blue line of $P = 99\%$ indicates that a concentration of at least 300 leads to 99% of the probability mass to be concentrated within a solid angle of $\alpha = 10^\circ$ around the mode. This is akin to the 3σ rule of the Gaussian distribution.

is:

$$\begin{aligned} P\left[\arccos(\mu^T n) < \alpha\right] &= \int_0^{2\pi}\int_0^{\alpha} Z(\tau)\exp(\tau\cos\phi)\sin\phi\,\mathrm{d}\phi\,\mathrm{d}\theta \\ &= 2\pi\int_0^{\alpha} Z(\tau)\exp(\tau\cos\phi)\sin\phi\,\mathrm{d}\phi \\ &= \frac{2\pi Z(\tau)}{\tau}\left(\exp\tau - \exp\left(\tau\cos\alpha\right)\right) \\ &= \frac{\exp\tau - \exp\left(\tau\cos\alpha\right)}{\exp\tau - \exp(-\tau)} \\ &= \frac{1-\exp\left(\tau(\cos\alpha - 1)\right)}{1-\exp(-2\tau)}\,. \end{aligned} \tag{24}$$

The cdf is shown in Fig. 5 to the left for different concentrations τ. The plot to the right shows the solid angle α as a function of concentration and probability P. The blue line for $P = 99\%$ shows the equivalent to the 3σ rule for the Gaussian distribution: for a concentration $\tau = 100$ 99% of the probability mass is within a solid angle of approximately $\alpha = 18^\circ$. To get to a probability mass of 99% inside a solid angle of 10° a concentration of $\tau = 300$ is needed. Such intuitions are useful when judging an inferred τ or choosing a fixed concentration.

5. Maximum Likelihood Estimate Parameters μ and τ in 3D

For ML estimation of the von-Mises-Fisher parameters it will be convenient to work in log scale. The log likelihood of a set of data $\{x_i\}_{i=1}^N$ is:

$$\log p(\{n_i\} \mid \mu, \tau) = \sum_i \log \text{vMF}(n_i \mid \mu, \tau) \tag{25}$$

$$= \tau\mu^T \sum_i n_i + N\log\tau - N\log\left(2\pi(\exp\tau - \exp(-\tau))\right)\,. \tag{26}$$

The ML estimate for μ can directly be read of: independent of the concentration τ the maximum with respect to the mode μ is attained if $\mu \in \mathbb{S}^2$ is directionally aligned with the sum over data vectors:

$$\mu^\star = \left\lceil \textstyle\sum_{i=1}^N n_i \right\rceil\,. \tag{27}$$

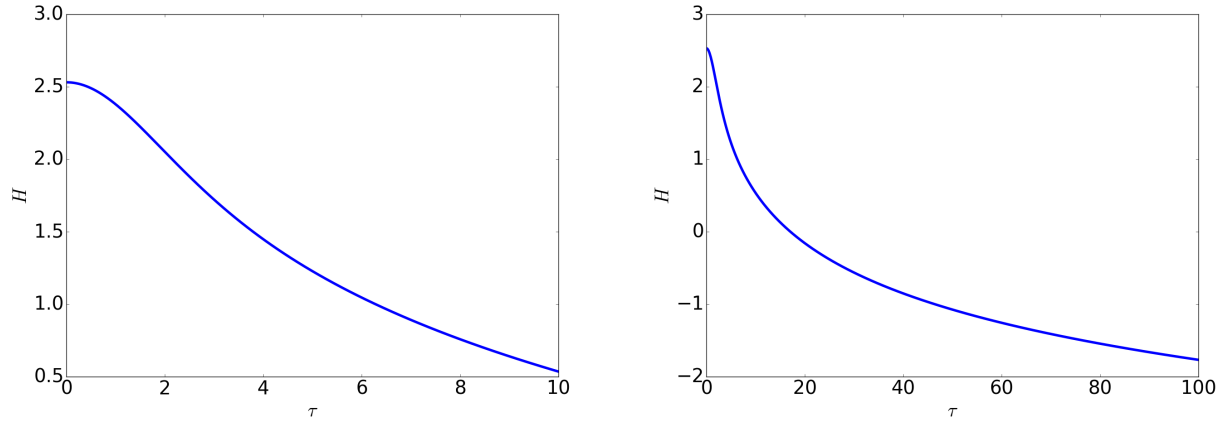


Figure 6: Entropy of the vMF distribution on $\mathbb{S}^2$ as a function of the concentration τ. The plot to the left is zoomed in to the range of $\tau \in [0, 10]$ whereas the plot to the right shows a larger range $\tau \in [0, 100]$.

To derive the maximum likelihood estimator for the concentration τ we set the derivative to 0:

$$\frac{\partial}{\partial \tau} \log p(\{n_i\} \mid \mu, \tau) = 0 \tag{28}$$

$$\tfrac{1}{N} \mu^T \textstyle\sum_i n_i + \tfrac{1}{\tau} - \tfrac{1+\exp(-2\tau))}{1-\exp(-2\tau)} = 0\,. \tag{29}$$

Since no closed form solution can be found we resort to Newton's method to efficiently obtain the ML estimate for the concentration τ. The thus obtained Extremum is indeed a maximum since the second derivative of the log likelihood is always negative:

$$\frac{\partial^2}{\partial \tau^2} \log p(\{n_i\} \mid \mu, \tau) = -\tfrac{1}{\tau^2} + \tfrac{4\exp(2\tau))}{(\exp(2\tau)-1)^2} < 0\,. \tag{30}$$

6. Entropy in 3D

The entropy of a von-Mises-Fisher distribution in 3D is computed as:

$$H = -\int_{x \in \mathbb{S}^2} \text{vMF}(x; \mu, \tau) \log \text{vMF}(x; \mu, \tau)\, \mathrm{d}x \tag{31}$$

$$= -\int_{x \in \mathbb{S}^2} \text{vMF}(x; \mu, \tau)(\log Z(\tau) + x^T \mu \tau)\, \mathrm{d}x \tag{32}$$

$$= -\log(Z(\tau)) - \tau Z(\tau) \int_{x \in \mathbb{S}^2} \exp(x^T \mu \tau) x^T \mu\, \mathrm{d}x \tag{33}$$

$$= -\log(Z(\tau)) - \tau Z(\tau) \int_0^{\pi} \int_0^{2\pi} \exp(\tau \cos\phi) \cos\phi \sin\phi\, \mathrm{d}\theta\, \mathrm{d}\phi \tag{34}$$

$$= -\log(Z(\tau)) - 2\pi\tau Z(\tau) \int_0^{\pi} \exp(\tau \cos\phi) \cos\phi \sin\phi\, \mathrm{d}\phi \tag{35}$$

$$= -\log(Z(\tau)) - 2\pi\tau Z(\tau) \frac{2\tau \cosh\tau - 2\sinh\tau}{\tau^2} \tag{36}$$

$$= -\log\left(\frac{\tau}{4\pi \sinh\tau}\right) - 2\pi \frac{\tau^2}{4\pi \sinh\tau} \frac{2\tau \cosh\tau - 2\sinh\tau}{\tau^2} \tag{37}$$

$$= -\log\left(\frac{\tau}{4\pi \sinh\tau}\right) - \frac{2\tau \cosh\tau - 2\sinh\tau}{2\sinh\tau} \tag{38}$$

$$= -\log\left(\frac{\tau}{4\pi \sinh\tau}\right) - \frac{\tau}{\tanh\tau} + 1\,, \tag{39}$$

where we have used $\mu = (0, 0, 1)$ without loss of generality (the integral and therefore the entropy is invariant to position of μ). At $\tau = 0$ the vMF distribution is uniform over the sphere. Hence its entropy is equivalent to the entropy of a uniform distribution over $\mathbb{S}^2$ which is $\log(4\pi) \approx= 2.53$, as can be verified in Fig. 6.

A. Appendix

A.1. Analysis of the Joint Prior for the von-Mises-Fisher Distribution

As introduced in Sec. 3, the joint prior of the vMF distribution is known up to proportionality as

$$p(\mu, \tau; \mu_0, a, b) \propto f(\tau, \mu; a, b, \mu_0) = \left(\frac{\tau}{\sinh \tau}\right)^a \exp\left(\tau b \mu^T \mu_0\right) \tag{40}$$

We will now characterize this distribution with a focus on the variation in τ to facilitate the implementation and theoretical justification of a slice sampler to sample from $p(\mu \mid \tau; \mu_0, a, b)$. Note that all analysis applies for the posterior distribution as well by using the posterior parameter a_N, b_N, and μ_N instead of a, b and μ_0. It will be convenient to work in log space:

$$\log f(\tau, \mu; a, b, \mu_0) = a \log \tau - a \log \sinh \tau + \tau b \mu^T \mu_0 \tag{41}$$

$$= a \log \tau + a \log 2 - a \log(1 - \exp(-2\tau)) + \tau\left(b \mu^T \mu_0 - a\right) \tag{42}$$

Keeping in mind that $0 < b < a$, the limit of the function as $\tau \to 0$ is 0 and as $\tau \to \infty$ is $-\infty$.

The derivative of $\log f(\tau)$ is

$$\frac{\partial}{\partial \tau} \log f(\tau) = \frac{a}{\tau} - \frac{2a \exp(-2\tau)}{1 - \exp(-2\tau)} + b \mu^T \mu_0 - a \tag{43}$$

Unfortunately once cannot solve for the maximum in closed form by setting the derivative to 0 (I have tried). The limits of this first derivative as $\tau \to 0$ is $b \mu^T \mu_0$ and as $\tau \to \infty$ is $b \mu^T \mu_0 - a$.

The second derivative of $\log f(\tau)$ is

$$\frac{\partial^2}{\partial \tau^2} \log f(\tau) = -\frac{a}{\tau^2} - a \frac{-4(1 - \exp(-2\tau)) \exp(-2\tau) - 4 \exp(-2\tau) \exp(-2\tau)}{(1 - \exp(-2\tau))^2} \tag{44}$$

$$= -\frac{a}{\tau^2} - a \frac{-4 \exp(-2\tau) + 4 \exp(-4\tau) - 4 \exp(-4\tau)}{(1 - \exp(-2\tau))^2} \tag{45}$$

$$= -\frac{a}{\tau^2} + \frac{4a \exp(-2\tau)}{(1 - \exp(-2\tau))^2} \tag{46}$$

$$= -\frac{a}{\tau^2} + \frac{4a \exp(-2\tau)}{1 - 2\exp(-2\tau) + \exp(-4\tau)} \tag{47}$$

The limit of this second derivative as $\tau \to 0$ is $-\frac{a}{3}$ and as $\tau \to \infty$ is 0. With $a > 0$ we can show that the second derivative is

always negative:

$$-\frac{a}{\tau^2}+\frac{4a\exp(-2\tau)}{1-2\exp(-2\tau)+\exp(-4\tau)}<0 \tag{48}$$

$$\frac{4\exp(-2\tau)}{1-2\exp(-2\tau)+\exp(-4\tau)}<\frac{1}{\tau^2} \tag{49}$$

$$4\tau^2\exp(-2\tau)<1-2\exp(-2\tau)+\exp(-4\tau) \tag{50}$$

$$(4\tau^2+2)\exp(-2\tau)<1+\exp(-4\tau) \tag{51}$$

$$2\tau^2+1<\frac{\exp(2\tau)+\exp(-2\tau)}{2} \tag{52}$$

$$2\tau^2+1<\cosh(2\tau) \tag{53}$$

$$2\tau^2+1<1+\frac{4\tau^2}{2}+\frac{16\tau^4}{24}+\cdots=\sum_{n=0}^{\infty}\frac{(2\tau)^{2n}}{(2n)!} \tag{54}$$

$$2\tau^2+1<1+2\tau^2+\frac{4\tau^4}{6}+\cdots=\sum_{n=0}^{\infty}\frac{(2\tau)^{2n}}{(2n)!} \tag{55}$$

$$0<\frac{4\tau^4}{6}+\cdots=\sum_{n=2}^{\infty}\frac{(2\tau)^{2n}}{(2n)!}\,. \tag{56}$$

The last statement is true since the infinite series is over strictly positive numbers because of the powers of even numbers $2n$. Therefore the second derivative is strictly negative with a limit of 0 for $\tau\to\infty$. This means that the first derivative is monotonically decreasing with a starting point (in the limit for $\tau\to 0$) of $b\mu^T\mu_0$ and an ending point at $b\mu^T\mu_0-a$ for $\tau\to\infty$. That in turn means that the first derivative has exactly one zero crossing (and hence the function one maximum) if $b\mu^T\mu_0\geq 0$ and none if $b\mu^T\mu_0<0$ (the largest function value is at 0). Therefore $\log f(\tau)$ is monotonically decreasing in the latter case and has a single maximum in the former.

The location of the maximum $\tau^\star$ cannot be computed in closed form, but we can use the Newton algorithm to obtain its location less than 10 iterations on average.

The locations of the zero-crossings of $g(\tau)=\log f(\tau)-\log(u)$ needed for the slice sampler are then obtained using Newtons method. Note that $g(\tau)$ has the same derivative as $\log f(\tau)$ derived in Eq. (43). The starting locations are set to $\tau_0^L=0.001\tau^\star$ for the left zero-crossing and to $\tau_0^R=1.5\tau^\star$ for the right zero-crossing. This ensures that Newton's method reliably converges to the desired zero crossing within a few iterations.

A.2. Normalizer of the Joint von-Mises-Fisher Prior for $D=3$ and $a=1$

We can derive a closed form normalizer for $a=1$, $0<b<a=1$ and $D=3$ dimensions:

$$Z(\mu_0,1,b)^{-1}=\int_0^\infty\int_{\mu\in\mathbb{S}^2}(2\pi)^{1/2}\,\frac{\tau\exp(\tau b\mu^T\mu_0)}{2\sinh\tau}\,\mathrm{d}\mu\,\mathrm{d}\tau \tag{57}$$

$$=\int_0^\infty 2^{-1/2}\pi^{1/2}\frac{\tau}{\sinh\tau}\int_{\mathbb{S}^2}\exp(\tau b\mu^T\mu_0)\,\mathrm{d}\mu\,\mathrm{d}\tau \tag{58}$$

$$=\int_0^\infty 2^{-1/2}\pi^{1/2}\frac{\tau}{\sinh\tau}Z^{-1}(\tau b)\,\mathrm{d}\tau \tag{59}$$

$$=\int_0^\infty 2^{-1/2}\pi^{1/2}\frac{\tau}{\sinh\tau}\frac{4\pi\sinh(\tau b)}{\tau b}\,\mathrm{d}\tau \tag{60}$$

$$=\frac{(2\pi)^{3/2}}{b}\int_0^\infty\frac{\sinh(\tau b)}{\sinh\tau}\,\mathrm{d}\tau \tag{61}$$

$$=\frac{2^{1/2}\pi^{5/2}}{b}\tan\left(\frac{b\pi}{2}\right),\forall-1<b<1 \tag{62}$$

where we have used that the integral over the vMF exponential term yields the normalizer of the vMF distribution.

A.3. Marginal Data Distribution of the von-Mises-Fisher Distribution

For $D = 3$ dimensions and $a = 1$ we can derive a closed form normalized probability density function for the marginal distribution of the data under the prior:

$$p(x_i; \mu_0, 1, b) = \int_0^\infty \int_{\mathbb{S}^2} \text{vMF}(x_i; \mu, \tau) p(\mu, \tau; \mu_0, 1, b) \, \mathrm{d}\mu \, \mathrm{d}\tau \tag{63}$$

$$= \int_0^\infty \int_{\mathbb{S}^2} \frac{\tau}{4\pi \sinh(\tau)} \frac{b\tau}{2\pi^2} \frac{\exp\left(\tau \mu^T (x_i + b\mu_0)\right)}{\tan\left(\frac{b\pi}{2}\right) \sinh(\tau)} \, \mathrm{d}\mu \, \mathrm{d}\tau \tag{64}$$

$$= \frac{b}{2^3 \pi^3 \tan\left(\frac{b\pi}{2}\right)} \int_0^\infty \frac{\tau^2}{\sinh^2(\tau)} \int_{\mathbb{S}^2} \exp\left(\tau \mu^T (x_i + b\mu_0)\right) \mathrm{d}\mu \, \mathrm{d}\tau \tag{65}$$

$$= \frac{b}{2^3 \pi^3 \tan\left(\frac{b\pi}{2}\right)} \int_0^\infty \frac{\tau^2}{\sinh^2(\tau)} Z^{-1} \left(\tau \|x_i + b\mu_0\|_2\right) \mathrm{d}\tau \tag{66}$$

$$= \frac{4\pi b}{2^3 \pi^3 \tan\left(\frac{b\pi}{2}\right)} \int_0^\infty \frac{\tau^2 \sinh(\tau \tilde{b})}{\tau \tilde{b} \sinh^2(\tau)} \, \mathrm{d}\tau \tag{67}$$

$$= \frac{b}{2\pi^2 \tilde{b} \tan\left(\frac{b\pi}{2}\right)} \int_0^\infty \frac{\tau \sinh(\tau \tilde{b})}{\sinh^2(\tau)} \, \mathrm{d}\tau \tag{68}$$

$$= \frac{b}{2\pi^2 \tilde{b} \tan\left(\frac{b\pi}{2}\right)} \frac{\pi(\tilde{b}\pi - \sin(\tilde{b}\pi))}{4 \sin^2(\frac{\tilde{b}\pi}{2})}, 0 < \tilde{b} < 2 \tag{69}$$

$$= \frac{b}{2^3 \tan\left(\frac{b\pi}{2}\right)} \frac{1 - \text{sinc}(\tilde{b}\pi)}{\sin^2(\frac{\tilde{b}\pi}{2})} \tag{70}$$

where we have used that the integrating over the vMF exponential term yields the inverse of the normalizer of the vMF distribution and that $0 < \tilde{b} < 2$ because $0 < b < a = 1$ which is imposed by the prior distributions properties.

References

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Books on Mathematics. Dover Publications, 1965.

[2] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway. Clustering on the unit hypersphere using von Mises-Fisher distributions. *JMLR*, 6(9), 2005.

[3] M. Bangert, P. Hennig, and U. Oelfke. Using an infinite von Mises-Fisher mixture model to cluster treatment beam directions in external radiation therapy. In *ICMLA*, pages 746–751. IEEE, 2010.

[4] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[6] S. Gopal and Y. Yang. von Mises-Fisher clustering models. In *ICML*, pages 154–162, 2014.

[7] R. M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[8] G. Nunez-Antonio and E. Gutiérrez-Pena. A Bayesian analysis of directional data using the von Mises-Fisher distribution. *Communications in Statistics-Simulation and Computation®*, 34(4):989–999, 2005.

[9] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 1999.

[10] J. Straub. *Nonparametric Directional Perception*. PhD thesis, Massachusetts Institute of Technology, May 2017.

[11] J. Straub, T. Campbell, J. P. How, and J. W. Fisher III. Small-variance nonparametric clustering on the hypersphere. In *CVPR*, 2015.

[12] G. Ulrich. Computer generation of distributions on the m-sphere. *Applied Statistics*, pages 158–163, 1984.