Diehe, T. (2015). A Note on the Kullback-Leibler Divergence for the von Mises-Fisher distribution. *arXiv.*

Peer reviewed version

[Link to publication record in Explore Bristol Research](http://www.bristol.ac.uk/pure/about/ebr-terms)
A Note on the Kullback-Leibler Divergence for the von Mises-Fisher distribution

T.R. Diethe
Department of Electrical and Electronic Engineering
University of Bristol
February 26, 2015

Abstract

We present a derivation of the Kullback Leibler (KL)-Divergence (also known as Relative Entropy) for the von Mises Fisher (VMF) Distribution in $d$-dimensions.

1 Introduction

The von Mises Fisher (VMF) Distribution (also known as the Langevin Distribution [8]) is a probability distribution on the $(d - 1)$-dimensional hypersphere $S^{d-1}$ in $\mathbb{R}^d$ [3]. If $d = 2$ the distribution reduces to the von Mises distribution on the circle, and if $d = 3$ it reduces to the Fisher distribution on a sphere. It was introduced by [3] and has been studied extensively by [6, 7]. The first Bayesian analysis was in [5] and recently it has been used for clustering on a hypersphere by [2].

Figure 1: Three sets of 1000 points sampled from three VMF distributions on the 3D sphere with $\kappa = 1$ (blue), $\kappa = 10$ (green) and $\kappa = 100$ (red), respectively. The mean directions are indicated with arrows.
2 Preliminaries

2.1 Definitions

We will use $\log(z)$ to denote the natural logarithm of $z$ throughout this article. Before continuing it will be useful to define the Gamma function $\Gamma(z)$,

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, \quad z \in \mathbb{C}, \text{Re}(z) > 0$$  \hspace{1cm} (1)

$$\Gamma(z) = (z-1)!, \quad z \in \mathbb{Z}^+$$  \hspace{1cm} (2)

and its relation, the incomplete Gamma function $\Gamma(z,s)$,

$$\Gamma(z,s) = (s-1)!e^{-x} \sum_{m=0}^{s-1} \frac{z^m}{m!}, \quad z \in \mathbb{Z}^+$$  \hspace{1cm} (3)

and the Modified Bessel Function of the First Kind $I_\alpha(z)$,

$$I_\alpha(z) = \sum_{m=0}^{\infty} \frac{(z/2)^{2m+\alpha}}{m!(m+\alpha+1)}$$  \hspace{1cm} (4)

which also has the following integral representations [1],

$$I_\alpha(z) = \frac{(z/2)^{\alpha}}{\sqrt{\pi} \Gamma(\alpha + 1/2)} \int_0^\pi e^{\pm z \cos \theta} \sin^{2\alpha} \theta d\theta, \quad (\alpha \in \mathbb{R})$$  \hspace{1cm} (5)

$$= \frac{(z/2)^{\alpha}}{\sqrt{\pi} \Gamma(\alpha + 1/2)} \int_{-1}^1 (1 - t^2)^{(\alpha-1/2)}e^{\pm zt} dt. \quad (\alpha \in \mathbb{R}, \alpha > -0.5)$$  \hspace{1cm} (6)

Also of interest is the logarithm of this quantity (using the second integral definition [6]),

$$\log(I_\alpha(z)) = \log \left[ \frac{(z/2)^{\alpha}}{\sqrt{\pi} \Gamma(\alpha + 1/2)} \int_{-1}^1 (1 - t^2)^{(\alpha-1/2)}e^{\pm zt} dt \right]$$

$$= \log \left[ \frac{(z/2)^{\alpha}}{\sqrt{\pi} \Gamma(\alpha + 1/2)} \right] + \log \left[ \int_{-1}^1 (1 - t^2)^{(\alpha-1/2)}e^{\pm zt} dt \right]$$

$$= \log \left( \frac{z}{2} \right)^{\alpha} - \log \sqrt{\pi} \Gamma(\alpha + 1/2) + \log \left[ \int_{-1}^1 (1 - t^2)^{(\alpha-1/2)}e^{\pm zt} dt \right].$$  \hspace{1cm} (7)

Note that the second term does not depend on $z$.

The Exponential Integral function $E_\alpha(z)$ is given by,

$$E_\alpha(z) = \int_1^\infty \frac{e^{-zt}}{t^\alpha} dt, \quad z^\alpha - 1 \Gamma(1 - \alpha, z).$$  \hspace{1cm} (8)

An identity that will be useful is,

$$\int_{-1}^1 (1 - t)^d e^{\kappa t} = -2^{d-1}E_{-d}(2\kappa)e^\kappa. \quad d > 0$$  \hspace{1cm} (9)
2.2 The von Mises Fisher (VMF) distribution

The probability density function (PDF) of the VMF distribution for a random d-dimensional unit vector $x$ is given by:

$$M_d(\mu, \kappa) = c_d(\kappa)e^{\kappa \mu'x}, \quad x \in S^{d-1},$$

where the normalisation constant $c_d(\kappa)$ is given by,

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}.$$  \hspace{1cm} (11)

The (non-symmetric) Kullback Leibler (KL)-Divergence from one probability distribution $q(x)$ to another probability distribution $p(x)$ is defined as,

$$KL(q(x)||p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)} \, dx,$$  \hspace{1cm} (12)

$$= \mathbb{E}_x \left[ \log \frac{q(x)}{p(x)} \right].$$  \hspace{1cm} (13)

Although this is general to any two distributions, we will assume that $p(x)$ is the “prior” distribution and $q(x)$ is the “posterior” distribution as commonly used in Bayesian analysis.

3 KL-Divergence for the VMF Distribution

3.1 General Case

We will assume that we have prior and posterior distributions defined over vectors $x \in \mathbb{R}^d$, $\|x\|_2 = 1$ as follows,

$$p(x) \sim M_d(\mu_p, \kappa_p),$$

$$q(x) \sim M_d(\mu_q, \kappa_q).$$  \hspace{1cm} (14)

We will now derive the KL-Divergence for two VMF distributions. The main problem in doing so will be the the normalisation constants $c_d(\kappa_p)$ and $c_d(\kappa_q)$.

**Theorem 3.1** For prior and posterior distributions as defined above over vectors $x \in \mathbb{R}^d$, $\|x\|_2 = 1$, $d < \infty$, $d$ odd, we have

$$KL(q(x)||p(x)) \leq \kappa_q - \kappa_p \mu_p' \mu_q + d^* \log(\kappa_q) + \sum_{m=1}^{d^*} \frac{\kappa_q^m}{m!}$$

$$- \left( \frac{d^2 - 2d + 1}{4} \right) \log(\kappa_p) + d^*(d^* + 1) \log d^* - d^* + 1$$

$$\hspace{1cm} (15)$$

**Proof** From (12), letting $d^* = \frac{d}{2} - 1$, $d^* = \frac{d-3}{2}$, and $d^* = \frac{d-1}{2}$, we have,

$$KL(q(x)||p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)} \, dx,$$

\hspace{1cm} (12)

\footnote{For even $d$ we can simply add a “null” dimension}
\[
= \int_x q(x) \left[ \log c_d (\kappa_q) e^{\kappa_q \mu_p x} - \log c_d (\kappa_p) e^{\kappa_p \mu_p x} \right] \, dx,
\]
\[
= \int_x q(x) \left[ \log c_d (\kappa_q) - \log c_d (\kappa_p) + \kappa_q \mu_p x - \kappa_p \mu_p x \right] \, dx,
\]
\[
= \int_x q(x) \left[ d^* \log (\kappa_q) - (d/2) \log (2 \pi) - \log I_{d^*} (\kappa_q) - d^* \log (\kappa_p) + \kappa_q \mu_p x - \kappa_p \mu_p x \right] \, dx,
\]
\[
= \int_x q(x) \left[ d^* \log \left( \frac{\kappa_q}{\kappa_p} \right) - \log I_{d^*} (\kappa_q) + \log I_{d^*} (\kappa_p) + \kappa_q \mu_p x - \kappa_p \mu_p x \right] \, dx
\]
\[
= \int_x q(x) \left[ d^* \log \left( \frac{\kappa_q}{\kappa_p} \right) + \kappa_q \mu_p x - \kappa_p \mu_p x \right.
\]
\[
- \log \left( \frac{\kappa_q}{2} \right)^{d^*} \log \sqrt{\Gamma} \left( d^* + \frac{1}{2} \right) - \log \int_{-1}^{1} \left( 1 - t^2 \right)^{d^*} e^{\pm \kappa_q \eta} \, dt
\]
\[
+ \log \left( \frac{\kappa_p}{2} \right)^{d^*} \log \sqrt{\Gamma} \left( d^* + \frac{1}{2} \right) + \log \int_{-1}^{1} \left( 1 - t^2 \right)^{d^*} e^{\pm \kappa_p \eta} \, dt \right] \, dx
\]
(Using (7))
\[
= \int_x q(x) \left[ d^* \log \left( \frac{\kappa_q}{\kappa_p} \right) + \kappa_q \mu_p x - \kappa_p \mu_p x \right.
\]
\[
- \log \int_{-1}^{1} \left( 1 - t^2 \right)^{d^*} e^{\pm \kappa_q \eta} \, dt + \log \int_{-1}^{1} \left( 1 - t^2 \right)^{d^*} e^{\pm \kappa_p \eta} \, dt \right] \, dx
\]
(Using (8))
\[
= \int_x q(x) \left[ \kappa_q \mu_p x - \kappa_p \mu_p x \right.
\]
\[
- \log \left[ -2 \frac{\kappa_q}{2} E_{-d^*}(2\kappa_q) e^{\kappa_q \eta} \right] + \log \left[ -2 \frac{\kappa_p}{2} E_{-d^*}(2\kappa_p) e^{\kappa_p \eta} \right] \right] \, dx
\]
\[
= \int_x q(x) \left[ \kappa_q \mu_p x - \kappa_p \mu_p x - \kappa_q + \kappa_p - \log \left[ E_{-d^*}(2\kappa_q) \right] + \log \left[ E_{-d^*}(2\kappa_p) \right] \right] \, dx
\]
(Using the definition of the Exponential Integral function (8))
\[
= \int_x q(x) \left[ \kappa_q (\mu_p x - 1) - \kappa_p (\mu_p x - 1) \right.
\]
\[
- \log \left( 2\kappa_q^{-d^*} \Gamma (d^*, 2\kappa_q) \right) + \log \left( 2\kappa_p^{-d^*} \Gamma (d^*, 2\kappa_p) \right) \right] \, dx
\]
(17)
\[
= \int_x q(x) \left[ \kappa_q (\mu_p x - 1) - \kappa_p (\mu_p x - 1) + d^* \log(2\kappa_q) - d^* \log(2\kappa_p) \right.
\]
\[
- \log \left( \Gamma (d^*, 2\kappa_q) \right) + \log \left( \Gamma (d^*, 2\kappa_p) \right) \right] \, dx
\]
(Using the definition of the Exponential Integral function (8))
\[
= \int_x q(x) \left[ \kappa_q (\mu_p x - 1) - \kappa_p (\mu_p x - 1) + d^* \log(2\kappa_q) - d^* \log(2\kappa_p) \right.
\]
\[
- \log \left( d^* e^{-\kappa_q} \sum_{m=0}^{d^*} \frac{\kappa_q^m}{m!} \right) + \log \left( d^* e^{-\kappa_p} \sum_{m=0}^{d^*} \frac{\kappa_p^m}{m!} \right) \right] \, dx
\]
(Using (3) and that \( a^* - 1 = d^e \))

\[
\begin{align*}
&= \int_x q(x) \left[ \kappa_q (\mu'_p x - 1) - \kappa_p (\mu'_p x - 1) + d^* \log(\kappa_q) - d^* \log(\kappa_p) + \kappa_q - \kappa_p \right. \\
&\quad - \log \left( \sum_{m=0}^{d^e} \frac{\kappa_q^m}{m!} \right) + \log \left( \sum_{m=0}^{d^e} \frac{\kappa_p^m}{m!} \right) \right] \, dx \\
&= \int_x q(x) \left[ \kappa_q \mu'_p x - \kappa_p \mu'_p x + d^* \log(\kappa_q) - d^* \log(\kappa_p) \\
&\quad - \log \left( \sum_{m=0}^{d^e} \frac{\kappa_q^m}{m!} \right) + \log \left( \sum_{m=0}^{d^e} \frac{\kappa_p^m}{m!} \right) \right] \, dx
\end{align*}
\]

Further simplifications:

\[
\begin{align*}
&\leq \int_x q(x) \left[ \kappa_q \mu'_p x - \kappa_p \mu'_p x + d^* \log(\kappa_q) - d^* \log(\kappa_p) \\
&\quad - \sum_{m=0}^{d^e} \log \left( \frac{\kappa_q^m}{m!} \right) + \sum_{m=0}^{d^e} \log \left( \frac{\kappa_p^m}{m!} \right) \right] \, dx
\end{align*}
\]

(by Jensen’s inequality)

\[
\begin{align*}
&= \int_x q(x) \left[ \kappa_q \mu'_p x - \kappa_p \mu'_p x + d^* \log(\kappa_q) - d^* \log(\kappa_p) \\
&\quad + \log \left( \sum_{m=0}^{d^e} \frac{\kappa_q^m}{m!} \right) - \sum_{m=0}^{d^e} (m \log(\kappa_p) - m \log m!) \right] \, dx \\
&\leq \int_x q(x) \left[ \kappa_q \mu'_p x - \kappa_p \mu'_p x + d^* \log(\kappa_q) - d^* \log(\kappa_p) \\
&\quad + \log \left( \sum_{m=0}^{d^e} \frac{\kappa_q^m}{m!} \right) - \sum_{m=1}^{d^e} (m \log(\kappa_p) - m \log m + m - 1) \right] \, dx
\end{align*}
\]

(20)

(21)

(22)

(Using \( n \log \frac{n}{e} + 1 \leq \log n! \leq (n + 1) \log \frac{n+1}{e} + 1 \))
\[
\int x q(x) \left[ \kappa q \mu_q^t x - \kappa p \mu_p^t x + d^* \log(\kappa_q) - d^* \log(\kappa_p) + \log \left( \sum_{m=0}^{d^p} \frac{\kappa_q^m}{m!} \right) \right.
\]
\[-d^p (d^p + 1) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \bigg] dx
\]
\[
= \int x q(x) \left[ \kappa q \mu_q^t x - \kappa p \mu_p^t x + d^* \log(\kappa_q) - d^* \log(\kappa_p) + \log \left( \sum_{m=0}^{d_q} \frac{\kappa_q^m}{m!} \right) \right.
\]-
\[\left. \left( \frac{(d - 3)^2}{4} + \frac{d - 3}{2} \right) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \right] dx
\]
\[
= \int x q(x) \left[ \kappa q \mu_q^t x - \kappa p \mu_p^t x + d^* \log(\kappa_q) + \log \left( \sum_{m=0}^{d_q} \frac{\kappa_q^m}{m!} \right) \right.
\]-
\[\left. \left( \frac{d^2 - 2d + 1}{4} \right) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \right] dx
\]
\[
\leq \int x q(x) \left[ \kappa q \mu_q^t x - \kappa p \mu_p^t x + d^* \log(\kappa_q) + \sum_{m=1}^{d_p} \frac{\kappa_q^m}{m!} \right.
\]-
\[\left. \left( \frac{d^2 - 2d + 1}{4} \right) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \right] dx
\]
(\text{using } n \geq \log(1 + n) \geq \frac{n}{1 + n}, \ (n > -1))
\[
= \int x q(x) \left[ \kappa q \mu_q^t x - \kappa p \mu_p^t x + d^* \log(\kappa_q) + \sum_{m=1}^{d_q} \frac{\kappa_q^m}{m!} \right.
\]-
\[\left. \left( \frac{d^2 - 2d + 1}{4} \right) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \right] dx
\]
\[= \kappa_q - \kappa q \mu_q^t \mu_q + d^* \log(\kappa_q) + \sum_{m=1}^{d_q} \frac{\kappa_q^m}{m!} \right.\]
\[-\left. \left( \frac{d^2 - 2d + 1}{4} \right) \log(\kappa_p) + d^p (d^p + 1) \log d^p - d^p + 1 \right) \]
(\text{as } \int x q(x) = 1, \text{ and } \mathbb{E}[x] = \mu_q, \text{ and } \mu_q^t \mu_q = 1) \]
(24)

The term \( \mu_q^t \mu_q \) can be seen as the cosine distance between the prior and posterior mean vectors. For \( 0 < \kappa_q < 1 \), the term \( \sum_{m=1}^{d} \frac{\kappa_q^m}{m!} \geq \kappa_q \). However for large \( \kappa_q \) and large \( d \) this term can grow very large.

**Special case: uniform prior**

Since the VMF distribution is defined on the \( S^{d-1} \), hypersphere, which is actually a specific case of a Stiefel manifold where \( r = 1 \) is the radius. The Stiefel
manifold has finite area,
\[ \tau(d, r) = \frac{2^r \pi^{\frac{r}{2}}}{\pi^{\frac{r}{2} + \frac{1}{2}} \prod_{j=1}^{r} \Gamma \left( \frac{r-j+1}{2} \right)}, \]  
(25)
and so,
\[ \tau(d, 1) = \frac{2\pi^{\frac{d}{2}}}{\Gamma \left( \frac{d}{2} \right)}, \]  
(26)
For the special case of the uniform prior (more precisely \( \lim_{\kappa \to 0} \)), the prior PDF reduces to,
\[ M_d(\mu, \kappa) = c_d(0)e^0 \]
\[ = \frac{\Gamma \left( \frac{d}{2} \right)}{2\pi^{\frac{d}{2}}}, \]  
(27)
which is simply one over the area on the manifold. This leads to a simpler form for the KL-divergence.

**Corollary 3.2** For prior and posterior distributions as defined above over vectors \( x \in \mathbb{R}^d, \|x\|_2 = 1, d < \infty \), we have
\[ \text{KL}(q(x)||p(x)) = \kappa_q - d^* \log 2 \]
(28)

**Proof**
\[
\text{KL}(q(x)||p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)} dx,
\]
\[ = \int_x q(x) \left[ \log c_d(\kappa_q) e^{\kappa_q \mu_q^T x} - \log c_d(0) \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x + \log c_d(\kappa_q) - \log \Gamma \left( \frac{d}{2} \right) + \log \left( 2\pi^{\frac{d}{2}} \right) \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x + \log c_d(\kappa_q) - \log(d^*)! + (d/2) \log (2\pi) \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x + d^* \log(\kappa_q) - (d/2) \log(2\pi) - \log I_d(\kappa_q) - \log(d^*)! \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x + d^* \log(\kappa_q) - \log I_d(\kappa_q) - \log(d^*)! \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x + d^* \log(\kappa_q) - \log \left( \kappa_q \frac{d}{2} \right) - \log(2\pi) \right] dx,
\]
\[ = \int_x q(x) \left[ \kappa_q \mu_q^T x - d^* \log 2 \right] dx,
\]
\[ = \kappa_q - d^* \log 2, \]  
(29)
For this special case, it can be seen that the dependence on the dimension is much more benign. This could prove useful for further computation (e.g. if the KL-divergence were to be used in a probably approximately correct (PAC)-Bayes bound [4]).

4 Conclusions

We have presented a derivation of the Kullback Leibler (KL)-divergence for the von Mises Fisher (VMF)-distribution, including the special case of a uniform prior over the hypersphere.

References

[1] Milton Abramowitz, Irene A Stegun, et al. Handbook of mathematical functions, volume 1. Dover New York, 1972.

[2] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. In Journal of Machine Learning Research, pages 1345–1382, 2005.

[3] Ronald Fisher. Dispersion on a sphere. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 217, pages 295–305. The Royal Society, 1953.

[4] John Langford. Tutorial on practical prediction theory for classification. In Journal of machine learning research, pages 273–306, 2005.

[5] Kanti V Mardia and SAM El-Atoum. Bayesian inference for the von mises-fisher distribution. Biometrika, 63(1):203–206, 1976.

[6] Kantilal Varichand Mardia. Statistics of directional data. Academic Press, 2014.

[7] KV Mardia and PJ Zemroch. Algorithm as 86: The von mises distribution function. Applied Statistics, pages 268–272, 1975.

[8] Yoko Watanori et al. Statistical inference of langevin distribution for directional data. Hiroshima Mathematical Journal, 26(1):25–74, 1996.