Kullback–Leibler Divergence and Moment Matching for Hyperspherical Probability Distributions

Gerhard Kurz, **Florian Pfaff**, and **Uwe D. Hanebeck**
Intelligent Sensor-Actuator-Systems Laboratory (ISAS)
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
gerhard.kurz@kit.edu, florian.pfaff@kit.edu, uwe.hanebeck@ieee.org

***Abstract*—When approximating one probability density with another density, it is desirable to minimize the information loss of the approximation as quantified by, e.g., the Kullback–Leibler divergence (KLD). It has been known for some time that in the case of the Gaussian distribution, matching the first two moments of the original density yields the optimal approximation in terms of minimizing the KLD. In this paper, we will show that a similar property can be proven for certain hyperspherical probability distributions, namely the von Mises–Fisher and the Watson distribution. This result has profound implications for moment-based filtering on the unit hypersphere as it shows that moment-based approaches are optimal in the information-theoretic sense.**

***Keywords*—von Mises–Fisher distribution, Watson distribution, parameter estimation**

I. Introduction

In many practical scenarios, it is necessary to approximate some complicated density $p(\cdot)$ with a simpler density $q(\cdot)$ that is more convenient to use. For example, a Gaussian mixture might be approximated with a single Gaussian component. In order to find a good approximation, we desire to minimize the information loss, e.g., the Kullback–Leibler divergence (KLD) [1] between the original and the approximating density. The KLD between two probability densities $p(\cdot)$ and $q(\cdot)$ on a domain D is defined as

$$\mathrm{KLD}(p||q) = \int_D p(\underline{x}) \log\left(\frac{p(\underline{x})}{q(\underline{x})}\right) \,\mathrm{d}\underline{x} \ .$$

This integral quantifies the information loss when approximating $p(\cdot)$ with $q(\cdot)$. The KLD can be rewritten as

$$\mathrm{KLD}(p||q) = \underbrace{-\int_D p(\underline{x}) \log\left(q(\underline{x})\right) \,\mathrm{d}\underline{x}}_{\text{cross entropy}} - \underbrace{\int_D -p(\underline{x}) \log(p(\underline{x})) \,\mathrm{d}x}_{\text{entropy of } p} \ ,$$

i.e., the difference of the cross entropy between $p(\cdot)$ and $q(\cdot)$ and the entropy of $p(\cdot)$ (see [2, Sec. 2.8.2]). As we assume $p(\cdot)$ to be given, minimizing the KLD is equivalent to minimizing the cross entropy.

It has been known for a long time that the Gaussian density possesses an interesting and useful property. If we approximate an arbitrary density on $\mathbb{R}^n$ with a Gaussian density, the parameters that minimize the KLD are exactly the same as the parameters obtained by setting the first two moments of the approximating (Gaussian) density to the first two moments of the given density (see, e.g., [3, Sec. 2]). Setting the parameters in such a way that the moments are equal is called *moment matching*. Thus, approximating the true density on $\mathbb{R}^n$ with a Gaussian by means of moment matching is justified in the sense that the information loss is minimized.

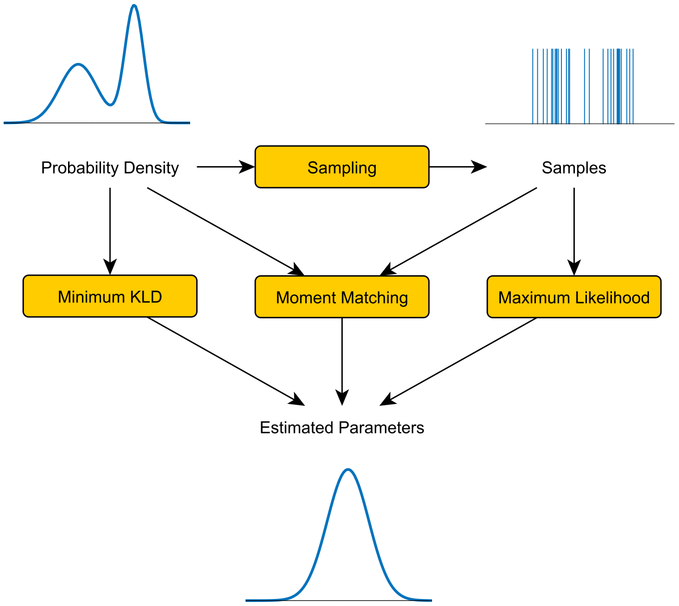


Fig. 1: Parameter estimation from a probability density or a set of samples using different methods. Depending on the original distribution, the different methods may or may not lead to the same result. We can also approximate a continuous probability density by an intermediate sample representation. In this case, we consider the limit for an infinite number of samples.

Beyond the Gaussian, there are other densities for which such a result can be obtained. In particular, Herbrich [4] showed a similar property for certain exponential densities on $\mathbb{R}^n$. It is also possible to show related properties for distributions defined on manifolds other than $\mathbb{R}^n$. Chiuso et al. [5, p. 95] proved that the von Mises–Fisher distribution on the unit sphere $S^2 \subset \mathbb{R}^3$ also has the property that moment matching (using the so-called mean result vector instead of traditional power moments) yields the same result as minimizing the KLD. Furthermore, we showed that the von Mises distribution on the unit circle also has this property if trigonometric moments are used [6]. In the following, we show a closely related property for certain hyperspherical probability distributions.

There are various applications of probability distributions on the sphere or hypersphere in numerous fields. Early uses of hyperspherical distributions can be found in geology [7], [8], [9] because they naturally appear when considering quantities such as the orientation of rock formations or paleomagnetic fields. In recent years, these approaches have found their way into many fields and are not limited to descriptive statistics, but are also used for estimation and filtering. For example, the task of tracking objects using omnidirectional cameras [10] can naturally be mapped to a spherical estimation problem. Thus, it even becomes possible to consider multiple target tracking on the surface of the unit sphere [11]. The orientation of crystalline structures including the occurring symmetries can also be elegantly represented using spherical distributions [12]. In the field of signal processing, hyperspherical distributions have been used for multiple speaker tracking [13] and speaker clustering [14]. Further applications include quaternion-based orientation estimation [15], [16], protein structure modeling in molecular biology [17], machine learning [18], [19], and neuroscience [20].

The estimation of a density's parameters based on a set of samples is a closely related problem to the approximation of a continuous density. In this case, moment matching and maximum likelihood estimation (MLE) can be used. For some densities, e.g., Gaussian densities, it can be shown that both approaches always yield the same parameters. If a continous density is to be approximated (and not a set of samples), MLE can only be used indirectly. MLE is not immediately applicable for approximation of continuous densities because the likelihood of obtaining a particular set of samples is considered. However, the original density can be sampled stochastically and those samples can be used for MLE. If the parameters converge to some fixed result as the number of samples approaches infinity, we can obtain an MLE-based approximation this way. This result may (or may not, depending on the density) be identical to the results obtained by moment matching and the minimizing the KLD. This connection is illustrated in Fig. 1.

The contribution of this paper can be summarized as follows. We consider the von Mises–Fisher distribution and the Watson distribution on the unit hypersphere S^{d-1} and show that they both fulfill the property that matching a suitable moment (the mean resultant vector or the covariance matrix, respectively) minimizes the KLD. In both cases, we can also show a close connection to MLE. In the case of parameter estimation based on a set of samples, the MLE coincides with the moment matching estimator, which in turn is equivalent to the minimum KLD estimator.

Our novel results have significant implications for filtering algorithms based on von Mises–Fisher or Watson distributions, e.g., [21], [5]. In particular, they serve as a justification for moment-based filters by proving that approximations that use moment matching are not just ad-hoc methods but provide an optimal approximation in the sense that the information loss is minimized. Furthermore, our results are of interest for the research area of minimum divergence filtering [22], [23], where filtering algorithms are constructed to minimize the Kullback–Leibler divergence rather than, say, the mean squared error (MSE).

II. von Mises–Fisher Distribution

The von Mises–Fisher (VMF) distribution is a probability distribution on the unit hypersphere $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\} \subset \mathbb{R}^d$ [24]. It includes the von Mises distribution [25] on the unit circle as a special case for $d = 2$. The VMF distribution is sometimes also referred to as the Langevin distribution [26].

Definition 1 (von Mises–Fisher Distribution) *The von Mises–Fisher distribution is given by the probability density function*

$$q(\underline{x}) = c_d(\kappa) \exp(\kappa \underline{\mu}^T \underline{x}) \ ,$$

where $\underline{x} \in S^{d-1}$, $\underline{\mu} \in S^{d-1}$, and $\kappa \geq 0$. It can be shown that the normalization constant $c_d(\kappa)$ is given by

$$c_d(\kappa) = \left(\int_{S^{d-1}} \exp(\kappa \underline{\mu}^T \underline{x}) \, \mathrm{d}\underline{x} \right)^{-1} = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \ .$$

The term $I_{d/2-1}(\kappa)$ refers to the modified Bessel function of order $d/2-1$ evaluated at κ [27, Sec. 9.6].

Intuitively, $\underline{\mu}$ represents the predominant direction on the hypersphere and κ determines the concentration around that direction. Examples of the VMF density are depicted in Fig. 2. In the three-dimensional case ($d = 3$), the normalization constant can be rewritten as

$$c_3(\kappa) = \frac{\kappa}{4\pi \sinh(\kappa)} \ . \tag{1}$$

The mean resultant vector of an arbitrary spherical density $f(\cdot)$ is given by [28, Sec. 4.2.2]

$$\underline{m} = \mathbb{E}(\underline{x}) = \int_{S^{d-1}} \underline{x} \cdot f(\underline{x}) \, \mathrm{d}\underline{x} \ .$$

If $f(\cdot)$ is a VMF distribution, we obtain

$$\begin{aligned} \underline{m} &= \int_{S^{d-1}} \underline{x} \cdot c_d(\kappa) \exp(\kappa \underline{\mu}^T \underline{x}) \, \mathrm{d}\underline{x} \\ &= \underline{\mu} \cdot A_d(\kappa) \end{aligned}$$

where

$$A_d(\kappa) = \frac{I_{d/2-1}(\kappa)}{I_{d/2}(\kappa)} \ .$$

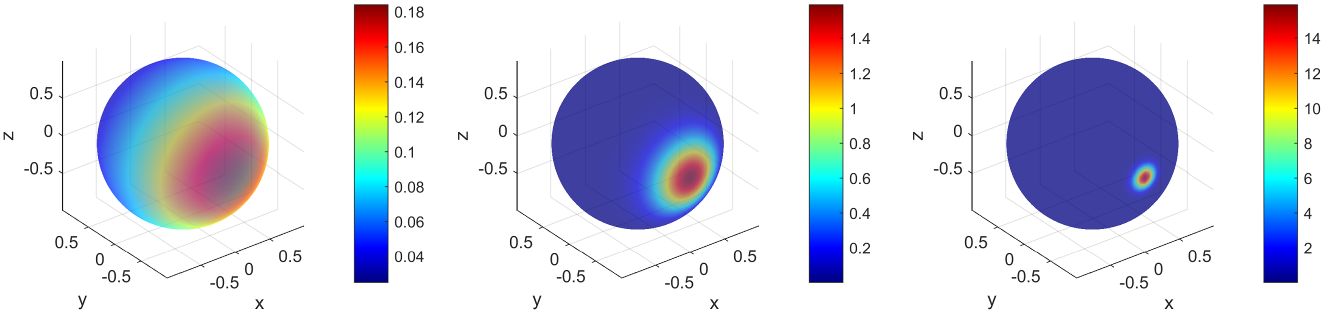
We give a proof of this property in Appendix A.

The moment-based parameter estimator is identical to the MLE method [29], [19] for parameter estimation and is given by the formulas

$$\begin{aligned} \underline{\mu} &= \frac{\underline{m}}{\|\underline{m}\|} \ , \\ \kappa &= A_d^{-1}(\|\underline{m}\|) \ , \end{aligned}$$

where $A_d^{-1}(\cdot)$ refers to the inverse function of $A_d(\cdot)$ and $\underline{m} \neq \underline{0}$ is the mean resultant vector of the spherical density or the sample set, respectively. In the case of $\underline{m} = \underline{0}$, $\underline{\mu}$ is an arbitrary unit vector and $\kappa = 0$.

We now formulate the main theorem regarding the connection between moment matching and the KLD in the case of a VMF distribution. The special case for S^2 has previously been shown by Chiuso et al. [5]. The main difference in the general case is that the normalization constant is significantly easier in 3D, as can be seen in (1).



(a) $\kappa = 1$.

(b) $\kappa = 10$.

(c) $\kappa = 100$.

Fig. 2: Densities of von Mises–Fisher distributions on the unit sphere S^2 in three dimensions. We use the location parameter $\underline{\mu} = [0, -1, 0]^T$ and different concentrations κ.

Theorem 1 *Consider an arbitrary probability density $p(\underline{x})$ on the unit hypersphere $S^{d-1} \subset \mathbb{R}^d$. Assume that $q(\underline{x}; \underline{\mu}, \kappa)$ follows a von Mises–Fisher distribution with parameters $\underline{\mu}$ and κ, i.e.,*

$$q(\underline{x}) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \exp(\kappa \underline{\mu}^T \underline{x}) \ .$$

Then

$$\arg\min_{\underline{\mu},\kappa} \mathit{KLD}(p || q(\underline{x}; \underline{\mu}, \kappa))$$

yields the same result as matching the mean resultant vector of $q(\underline{x})$ with that of $p(\underline{x})$.

Proof:

We can rewrite the KLD as

$$\begin{aligned}
&\mathrm{KLD}(p||q) \\
&= \int_{S^{d-1}} p(\underline{x}) \log \left(\frac{p(\underline{x})}{q(\underline{x}; \underline{\mu}, \kappa)} \right) \mathrm{d}\underline{x} \\
&= \int_{S^{d-1}} p(\underline{x}) \log p(\underline{x}) \,\mathrm{d}\underline{x} - \log \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \\
&\quad - \kappa \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \,\mathrm{d}\underline{x} \ .
\end{aligned}$$

As mentioned in the introduction, the first term is independent of $q(\cdot)$ and can be disregarded. In order to introduce the constraint $||\underline{\mu}|| = 1$ while minimizing the KLD, we use the Lagrange multiplier method. Thus, we can minimize the KLD by maximizing the function

$$\begin{aligned}
g(\underline{\mu}, \kappa, \lambda) := \log \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} + \kappa \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \,\mathrm{d}\underline{x} \\
+ \lambda (1 - \sum_{k=1}^{d} \mu_k^2)
\end{aligned}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier.

Now, we consider the partial derivatives with respect to μ_j for $j = 1, \dots, d$ and set them to zero

$$\begin{aligned}
\frac{\partial g}{\partial \mu_j} &= \kappa \int_{S^{d-1}} p(\underline{x}) x_j \,\mathrm{d}\underline{x} - 2\lambda \mu_j \overset{!}{=} 0 \\
\Leftrightarrow \mu_j &\overset{!}{=} \frac{\kappa}{2\lambda} \int_{S^{d-1}} p(\underline{x}) x_j \,\mathrm{d}\underline{x} \ ,
\end{aligned}$$

for $\lambda \neq 0$. Thus, we have

$$\mu_j \propto \int_{S^{d-1}} p(\underline{x}) x_j \,\mathrm{d}\underline{x} \ .$$

The second derivative is given by

$$\frac{\partial^2 g}{(\partial \mu_j)^2} = -2\lambda \ ,$$

i.e., we have a maximum for $\lambda > 0$ and a minimum for $\lambda < 0$. Considering the derivative of $g(\cdot)$ with respect to λ and setting it to zero yields

$$\frac{\partial g}{\partial \lambda} = 1 - \sum_{k=1}^{d} \mu_k^2 \overset{!}{=} 0 \ .$$

Thus, we obtain $\underline{\mu}$ as

$$\underline{\mu} = \frac{\underline{r}}{\|\underline{r}\|}, \quad \underline{r} := \int_{S^{d-1}} p(\underline{x}) \underline{x} \,\mathrm{d}\underline{x}$$

if $\|\underline{r}\| \neq 0$. If $\|\underline{r}\| = 0$, $g(\cdot)$ is independent of $\underline{\mu}$ and any unit vector can be chosen. In this case, we have $\kappa = 0$ and obtain a uniform distribution on the unit hypersphere.

For $\kappa > 0$, the derivative of $g(\cdot)$ with respect to κ is independent of λ and can be calculated as

$$\begin{aligned}
\frac{\partial g}{\partial \kappa} &= \frac{(2\pi)^{d/2} I_{d/2-1}(\kappa)}{\kappa^{d/2-1}} \cdot \frac{\partial}{\partial \kappa} \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \\
&\quad + \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \,\mathrm{d}\underline{x} \\
&= \frac{1}{\kappa^{d/2-1}} \cdot \frac{(d/2-1)\kappa^{d/2-2} I_{d/2-1}(\kappa) - \kappa^{d/2-1} \frac{\partial}{\partial \kappa} I_{d/2-1}(\kappa)}{I_{d/2-1}(\kappa)} \\
&\quad + \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \,\mathrm{d}\underline{x}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(d/2-1)\kappa^{-1} I_{d/2-1}(\kappa) - \frac{\partial}{\partial\kappa} I_{d/2-1}(\kappa)}{I_{d/2-1}(\kappa)} \\
&\quad + \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \, \mathrm{d}\underline{x} \\
&= \frac{(d/2-1)\kappa^{-1} I_{d/2-1}(\kappa) - ((d/2-1)\kappa^{-1} I_{d/2-1}(\kappa) + I_{d/2}(\kappa))}{I_{d/2-1}(\kappa)} \\
&\quad + \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \, \mathrm{d}\underline{x} \\
&= \frac{-I_{d/2}(\kappa))}{I_{d/2-1}(\kappa)} + \int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \, \mathrm{d}\underline{x}
\end{aligned}
$$

where we use the identity [30, eq. (A.8)] to compute the derivative of the Bessel function. Setting the derivative to zero yields

$$
\underbrace{\frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}}_{=:A_d(\kappa)} = \underbrace{\int_{S^{d-1}} p(\underline{x}) \underline{\mu}^T \underline{x} \, \mathrm{d}\underline{x}}_{=:R} \ .
$$

Thus, we have $\kappa = A_d^{-1}(R)$. Note that $R = \|r\|$ holds.

The second derivative with respect to κ

$$
\frac{\partial^2 g}{(\partial \kappa)^2} = -A_d'(\kappa)
$$

is smaller than zero for all $\kappa > 0$ as shown by [31, p. 242]. As a result, the derived values for $\underline{\mu}$ and κ constitute a maximum of $g(\cdot)$, and hence, minimize the KLD between $p(\cdot)$ and $q(\cdot)$. ■

This proof also shows the desired property for the circular case, i.e., $d = 2$. The VMF density on the unit circle can be parameterized using $\underline{x} = [\cos(\phi), \sin(\phi)]^T$ and $\underline{\mu} = [\cos(\nu), \sin(\nu)]^T$ according to

$$
\begin{aligned}
f(\underline{x}) &= f([\cos(\phi), \sin(\phi)]^T) \\
&= c_2(\kappa) \exp(\kappa [\cos(\nu), \sin(\nu)][\cos(\phi), \sin(\phi)]^T) \\
&= c_2(\kappa) \exp(\kappa (\cos(\nu)\cos(\phi) + \sin(\nu)\sin(\phi))) \\
&= c_2(\kappa) \exp(\kappa (\cos(\phi - \nu))) \ ,
\end{aligned}
$$

which is a von Mises density. Thus, the proof for the VMF distribution is also a generalization of the proof for the von Mises distribution given in [6].

III. Watson Distribution

In this section, we will focus on the Watson distribution [8], [9], which is closely related to the VMF distribution. The essential difference is the inclusion of a square in the exponent as seen in the following definition. The Watson distribution constitutes a special case of the Bingham distribution [7] with rotational symmetry.

Definition 2 (Watson Distribution) *The Watson distribution on S^{d-1} is given by the pdf*

$$
q(\underline{x}; \underline{\mu}, \kappa) = c_d(\kappa) \cdot \exp(\kappa (\underline{\mu}^T \underline{x})^2) \ ,
$$

where $\underline{x} \in S^{d-1}$, $\underline{\mu} \in S^{d-1}$ and $\kappa \in \mathbb{R}$. The normalization constant $c_d(\kappa)$ can be written as

$$
c_d(\kappa) = \frac{\Gamma(d/2)}{2\pi^{d/2} M(1/2, d/2, \kappa)} \ ,
$$

where $M(\cdot,\cdot,\cdot)$ is a confluent hypergeometric function known as Kummer's function [27, Sec. 13.1].

Kummer's function $M(\cdot,\cdot,\cdot)$ is given by the power series representation [30, eq. (A.18)]

$$
M(a, b, \kappa) = \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b)}{\Gamma(a)\Gamma(b+n)} \frac{\kappa^n}{n!} \ .
$$

Some examples of Watson densities are shown in Fig. 3. It can be seen that the density is antipodally symmetric, i.e., $f(\underline{x}) = f(-\underline{x})$. For $\kappa > 0$, the density is concentrated around the modes at $\underline{\mu}$ and $-\underline{\mu}$. As κ approaches zero, the density becomes less peaked and approaches a uniform distribution. For negative κ, the density is concentrated around the hyperplane through $\underline{0}$ that is perpendicular to $\underline{\mu}$.

As a result of the antipodal symmetry, it always holds that $\mathbb{E}(\underline{x}) = \underline{0}$. Consequently, the mean resultant vector of a Watson distribution does not contain any useful information. However, its second moment $\mathbf{C} = \mathbb{E}(\underline{x}\underline{x}^T)$ can be used for moment matching. Due to $\mathbb{E}(\underline{x}) = \underline{0}$ the second moment coincides with the second central moment, i.e., the covariance matrix. A derivation of this covariance matrix can be found in Appendix B.

The moment-based estimator can be found by inverting the covariance formula (see Appendix C) and is identical to the MLE estimator given in [32], [18]. To compute the MLE, we consider the smallest and largest eigenvalues of $\mathbf{C}$ with the corresponding eigenvectors. If $\kappa > 0$, $\underline{\mu}$ is the eigenvector corresponding to the largest eigenvalue. Otherwise, $\underline{\mu}$ is the eigenvector corresponding to the smallest eigenvalue. Following [30, eqs. (A.20), (A.22)], we define

$$
\begin{aligned}
D_d(\kappa) &= \frac{M'(1/2, d/2, \kappa)}{M(1/2, d/2, \kappa)} \qquad (2) \\
&= \frac{M(3/2, d/2+1, \kappa)}{d \cdot M(1/2, d/2, \kappa)} \ .
\end{aligned}
$$

We observe that the derivative of $M(\cdot,\cdot,\cdot)$ can easily be eliminated. The value of κ is then obtained by solving the equation

$$
D_d(\kappa) = \underline{\mu}^T \mathbf{C} \underline{\mu}
$$

for both possible values for $\underline{\mu}$ are then selecting the result that yields the larger likelihood. Note that the evaluation of the inverse $D_d^{-1}(\cdot)$ is only possible numerically as described in [32].

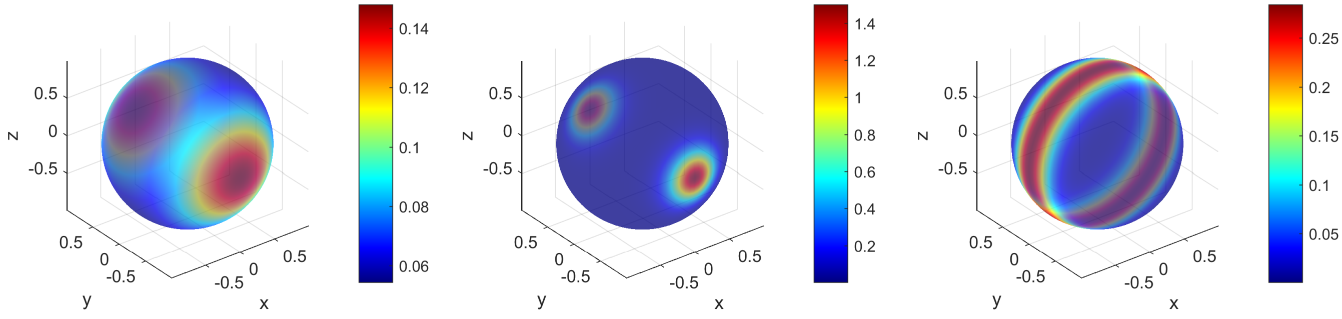
Theorem 2 *Consider an arbitrary probability density $p(\underline{x})$ on the unit hypersphere $S^{d-1} \subset \mathbb{R}^d$. Assume that $q(\underline{x}; \underline{\mu}, \kappa)$ follows a Watson distribution with parameters $\underline{\mu}$ and κ. Then*

$$
\arg\min_{\underline{\mu},\kappa} \mathit{KLD}(p || q(\underline{x}; \underline{\mu}, \kappa))
$$

yields the same result as matching the second moment of $q(\underline{x})$ to the second moment

$$
\mathbf{C} = \int_{S^{d-1}} \underline{x}\underline{x}^T \cdot p(\underline{x}) \, \mathrm{d}\underline{x}
$$

of $p(\underline{x})$.



(a) $\kappa = 1$.

(b) $\kappa = 10$.

(c) $\kappa = -10$.

Fig. 3: Densities of Watson distributions on the unit sphere S^2 in three dimensions. We use the location parameter $\underline{\mu} = [0, -1, 0]^T$ and different concentrations κ.

Proof: First, we rewrite the KLD according to

$$\begin{aligned}
&\mathrm{KLD}(p||q)\\
&= \int_{S^{d-1}} p(\underline{x}) \log\left(\frac{p(\underline{x})}{q(\underline{x};\underline{\mu},\kappa)}\right) \mathrm{d}\underline{x}\\
&= \int_{S^{d-1}} p(\underline{x}) \log p(\underline{x})\, \mathrm{d}\underline{x} - \int_{S^{d-1}} p(\underline{x}) \log q(\underline{x};\underline{\mu},\kappa)\, \mathrm{d}\underline{x}\\
&= \int_{S^{d-1}} p(\underline{x}) \log p(\underline{x})\, \mathrm{d}\underline{x} - \log c_d(\kappa) - \kappa \int_{S^{d-1}} p(\underline{x}) (\underline{\mu}^T \underline{x})^2\, \mathrm{d}\underline{x}\\
&= \int_{S^{d-1}} p(\underline{x}) \log p(\underline{x})\, \mathrm{d}\underline{x} - \log \Gamma(d/2) + \log(2\pi^{d/2})\\
&\quad + \log(M(1/2, d/2, \kappa)) - \kappa \int_{S^{d-1}} p(\underline{x}) (\underline{\mu}^T \underline{x})^2\, \mathrm{d}\underline{x} \ .
\end{aligned}$$

For fixed κ, we drop all terms independent of $\underline{\mu}$ and obtain

$$\mathrm{KLD}(p||q) = -\kappa \underline{\mu}^T \underbrace{\int_{S^{d-1}} p(\underline{x}) \underline{x}\underline{x}^T\, \mathrm{d}\underline{x}}_{=\mathbf{C}} \underline{\mu} + \mathrm{const} \ .$$

Because $||\underline{\mu}|| = 1$, we can see that $\underline{\mu}$ is the unit eigenvector of $\mathbf{C}$ corresponding to the largest eigenvalue (for $\kappa \geq 0$) or to the smallest eigenvalue (for $\kappa < 0$). Similar to [32], we calculate the result for both possible eigenvectors and later evaluate the KLD to find the optimal solution.

To obtain the optimal value of κ, we consider the derivative with respect to κ

$$\frac{\partial \mathrm{KLD}(p||q)}{\partial \kappa} = \frac{M'(1/2, d/2, \kappa)}{M(1/2, d/2, \kappa)} - \underline{\mu}^T \mathbf{C} \underline{\mu} \ .$$

Setting the partial derivative to zero and solving for κ yields $\kappa = D_d^{-1}(\underline{\mu}^T \mathbf{C} \underline{\mu})$. According to [32], this solution constitutes a minimum. ■

IV. Conclusion

We have shown a fundamental property of two hyperspherical probability distributions, the von Mises–Fisher distribution and the Watson distribution. When approximating an arbitrary hyperspherical distribution with one of these two distributions, the result obtained by moment matching is also optimal in the sense that the Kullback–Leibler divergence between the original density and its approximation is minimized. Thus, moment matching minimizes the information loss as a result of the approximation.

The novel results have interesting implications for estimation algorithms based on hyperspherical distributions. In particular, the use of moment-based approximations is now justified by an information-theoretic foundation. Thus, reliance on moment matching is not only motivated by its computational tractability, but also by the fact that it can be shown to be optimal in terms of minimizing the information loss.

Future work may include the investigation of other densities from directional statistics. It is expected that some of them can be shown to have similar properties, whereas for others counterexamples may be found.

Appendix A
Derivation of the Mean Resultant Vector of the VMF Distribution

In the following, we show the property

$$\begin{aligned}
\underline{m} &= \int_{S^{d-1}} \underline{x} \cdot c_d(\kappa) \exp(\kappa \underline{\mu}^T \underline{x})\, \mathrm{d}\underline{x}\\
&= \underline{\mu} \cdot A_d(\kappa) \ .
\end{aligned}$$

First, we consider the case of $\underline{\mu} = [1, 0, \dots, 0]^T$. For $2 \leq j \leq d$, we have

$$m_j = \int_{S^{d-1}} x_j \cdot c_d(\kappa) \exp(\kappa \mu_1 x_1)\, \mathrm{d}\underline{x} = 0$$

for reasons of symmetry of S^{d-1}. The first entry of $\underline{m}$ is derived according to

$$\begin{aligned}
m_1 &= \int_{S^{d-1}} x_1 \cdot c_d(\kappa) \exp(\kappa \mu_1 x_1)\, \mathrm{d}\underline{x}\\
&= c_d(\kappa) \int_{S^{d-1}} x_1 \exp(\kappa \mu_1 x_1)\, \mathrm{d}\underline{x}\\
&= \frac{c_d(\kappa)}{\mu_1} \int_{S^{d-1}} \frac{\partial}{\partial \kappa} \exp(\kappa \mu_1 x_1)\, \mathrm{d}\underline{x}\\
&= c_d(\kappa) \frac{\partial}{\partial \kappa} \int_{S^{d-1}} \exp(\kappa \mu_1 x_1)\, \mathrm{d}\underline{x}
\end{aligned}$$

$$= c_d(\kappa)\frac{\partial}{\partial\kappa}\frac{1}{c_d(\kappa)} \ ,$$

where we use the dominated convergence theorem to interchange differentiation and integration. Furthermore, we introduce the abbreviation $\hat{d} := \frac{d}{2} - 1$ and use [30, eq. (A.8)] to obtain

$$\begin{aligned}
&c_d(\kappa)\frac{\partial}{\partial\kappa}\frac{1}{c_d(\kappa)} \\
&= \frac{\kappa^{\hat d}}{(2\pi)^{\frac{d}{2}} I_{\hat d}(\kappa)}\frac{\partial}{\partial\kappa}\frac{(2\pi)^{\frac{d}{2}} I_{\hat d}(\kappa)}{\kappa^{\hat d}} \\
&= \frac{\kappa^{\hat d}}{I_{\hat d}(\kappa)}\frac{\partial}{\partial\kappa}\frac{I_{\hat d}(\kappa)}{\kappa^{\hat d}} \\
&= \frac{\kappa^{\hat d}}{I_{\hat d}(\kappa)}\frac{\left(\frac{\partial}{\partial\kappa} I_{\hat d}(\kappa)\right)\kappa^{\hat d} - \hat{d}\kappa^{\frac{d}{2}-2} I_{\hat d}(\kappa)}{(\kappa^{\hat d})^2} \\
&= \frac{\left(\frac{\partial}{\partial\kappa} I_{\hat d}(\kappa)\right)\kappa^{\hat d} - \hat{d}\kappa^{\frac{d}{2}-2} I_{\hat d}(\kappa)}{\kappa^{\hat d} I_{\hat d}(\kappa)} \\
&= \frac{(\hat{d} I_{\hat d}(\kappa)\kappa^{-1} + I_{\frac{d}{2}}(\kappa))\kappa^{\hat d} - \hat{d}\kappa^{\frac{d}{2}-2} I_{\hat d}(\kappa)}{\kappa^{\hat d} I_{\hat d}(\kappa)} \\
&= \frac{\hat{d} I_{\hat d}(\kappa)\kappa^{-1} + I_{\frac{d}{2}}(\kappa) - \hat{d}\kappa^{-1} I_{\hat d}(\kappa)}{I_{\hat d}(\kappa)} \\
&= \frac{I_{\frac{d}{2}}(\kappa)}{I_{\hat d}(\kappa)} = A_d(\kappa) \ .
\end{aligned}$$

For $\underline{\mu} \neq [1, 0, \ldots, 0]^T$, we consider an arbitrary rotation matrix $\mathbf{M}$ whose first column is $\underline{\mu}$. Thus, we have $\underline{\mu} = \mathbf{M} \cdot [1, 0, \ldots, 0]^T$ and $\underline{\mu}^T \mathbf{M} = [1, 0, \ldots, 0]$. We use integration by substitution with $\underline{x} = \mathbf{M}\underline{t}$ (which does not change the integration area and has volume correction term 1) to reduce the problem to the case of $\underline{\mu} = [1, 0, \ldots, 0]^T$, which yields

$$\begin{aligned}
\underline{m} &= \int_{S^{d-1}} \underline{x} \cdot c_d(\kappa)\exp(\kappa\underline{\mu}^T\underline{x})\,\mathrm{d}\underline{x} \\
&= \int_{S^{d-1}} \mathbf{M}\underline{t} \cdot c_d(\kappa)\exp(\kappa\underline{\mu}^T\mathbf{M}\underline{t})\,\mathrm{d}\underline{t} \\
&= \int_{S^{d-1}} \mathbf{M}\underline{t} \cdot c_d(\kappa)\exp(\kappa[1, 0, \ldots, 0]\underline{t})\,\mathrm{d}\underline{t} \\
&= \mathbf{M}\int_{S^{d-1}} \underline{t} \cdot c_d(\kappa)\exp(\kappa[1, 0, \ldots, 0]\underline{t})\,\mathrm{d}\underline{t} \\
&= \mathbf{M}[A_d(\kappa), 0, \ldots, 0]^T \\
&= \underline{\mu} \cdot A_d(\kappa) \ .
\end{aligned}$$

Appendix B
Derivation of the Covariance Matrix for the Watson Distribution

The following derivation can be seen as a special case of the covariance of a Bingham distribution [15, eq. (16)], [33, Sec. A.5]. However, the solution for the Bingham distribution is needlessly complicated as it relies on the matrix version of the confluent hypergeometric function.

First, we consider the case $\underline{\mu} = [1, 0, \ldots, 0]^T$. We have

$$\begin{aligned}
\mathbb{E}(\underline{x}\underline{x}^T) &= \int_{S^{d-1}} \underline{x}\underline{x}^T c_d(\kappa)\exp(\kappa(\underline{\mu}^T\underline{x})^2)\,\mathrm{d}\underline{x} \\
&= c_d(\kappa)\int_{S^{d-1}} \underline{x}\underline{x}^T \exp(\kappa x_1^2)\,\mathrm{d}\underline{x} \ .
\end{aligned}$$

The upper left entry of the matrix is found according to

$$\begin{aligned}
\mathbb{E}(x_1^2) &= c_d(\kappa)\int_{S^{d-1}} x_1^2\exp(\kappa x_1^2)\,\mathrm{d}\underline{x} \\
&= c_d(\kappa)\int_{S^{d-1}} \frac{\partial}{\partial\kappa}\exp(\kappa x_1^2)\,\mathrm{d}\underline{x} \\
&= c_d(\kappa)\frac{\partial}{\partial\kappa}\int_{S^{d-1}}\exp(\kappa x_1^2)\,\mathrm{d}\underline{x} \\
&= c_d(\kappa)\frac{\partial}{\partial\kappa}\frac{1}{c_d(\kappa)} \ ,
\end{aligned}$$

where we are able to exchange integration and differentiation due to the dominated convergence theorem. We further simplify

$$\begin{aligned}
&c_d(\kappa)\frac{\partial}{\partial\kappa}\frac{1}{c_d(\kappa)} \\
&= \frac{\Gamma(d/2)}{2\pi^{d/2}M(1/2, d/2, \kappa)}\frac{\partial}{\partial\kappa}\frac{2\pi^{d/2}M(1/2, d/2, \kappa)}{\Gamma(d/2)} \\
&= \frac{1}{M(1/2, d/2, \kappa)}\frac{\partial}{\partial\kappa}M(1/2, d/2, \kappa) \\
&= \frac{M(3/2, d/2+1, \kappa)}{d \cdot M(1/2, d/2, \kappa)} = D_d(\kappa),
\end{aligned}$$

where we use (2). Furthermore, it is easy to see that the off-diagonal entries are all zero because for $i \neq j$

$$\mathbb{E}(x_i \cdot x_j) = c_d(\kappa)\int_{S^{d-1}} x_i x_j \exp(\kappa x_1^2)\,\mathrm{d}\underline{x} = 0$$

due to symmetry of S^{d-1}. To obtain the remaining entries, we observe that

$$\operatorname{trace}\mathbb{E}(\underline{x}\underline{x}^T) = \mathbb{E}(\operatorname{trace}\underline{x}\underline{x}^T) = \mathbb{E}(\operatorname{trace}\underline{x}^T\underline{x}) = \mathbb{E}(1) = 1$$

because all $\underline{x}$ are unit vectors. This implies $\sum_{i=1}^d \mathbb{E}(x_i^2) = 1$. Furthermore, we have $\mathbb{E}(x_i \cdot x_i) = \mathbb{E}(x_j \cdot x_j)$ for all $2 \leq i, j \leq d$. Thus, it holds that

$$\mathbb{E}(x_i \cdot x_i) = \frac{1 - \mathbb{E}(x_1^2)}{d-1} = \frac{1 - c_d(\kappa)\frac{\partial}{\partial\kappa}\frac{1}{c_d(\kappa)}}{d-1} = \frac{1 - D_d(\kappa)}{d-1}$$

for $2 \leq i \leq d$.

To generalize the derivation to arbitrary $\underline{\mu}$, we use the same technique as in the case of the VMF distribution. Specifically, we consider a rotation matrix $\mathbf{M}$ whose first column is $\underline{\mu}$. A suitable matrix can be obtained using the QR decomposition (see [21], [34]). According to

$$\begin{aligned}
&\mathbb{E}(\underline{x}\underline{x}^T) \\
&= \int_{S^{d-1}} \underline{x}\underline{x}^T c_d(\kappa)\exp(\kappa(\underline{\mu}^T\underline{x})^2)\,\mathrm{d}\underline{x} \\
&= \int_{S^{d-1}} \mathbf{M}\underline{t}(\mathbf{M}\underline{t})^T c_d(\kappa)\exp(\kappa(\underline{\mu}^T\mathbf{M}\underline{t})^2)\,\mathrm{d}\underline{t} \\
&= \mathbf{M} \cdot \int_{S^{d-1}} \underline{t}\underline{t}^T c_d(\kappa)\exp(\kappa([1, 0, \ldots, 0]^T\underline{t})^2)\,\mathrm{d}\underline{t} \cdot \mathbf{M}^T
\end{aligned}$$

we can reduce the problem to the case with $\underline{\mu} = [1, 0, \dots, 0]^T$. Therefore, the complete result is given by

$$\mathbb{E}(\underline{x}\underline{x}^T) = \mathbf{M} \operatorname{diag}\left(D_d(\kappa), \frac{1-D_d(\kappa)}{d-1}, \dots, \frac{1-D_d(\kappa)}{d-1} \right) \mathbf{M}^T .$$

Appendix C
Derivation of the Moment-based Estimator for the Watson Distribution

We want to obtain the parameters $\underline{\mu}$ and κ from a given second moment matrix $\mathbf{C} = \mathbb{E}(\underline{x}\underline{x}^T)$. As $\mathbf{C}$ is symmetric positive definite, we can compute the eigenvalue decomposition $\mathbf{C} = \mathbf{M}\mathbf{\Sigma}\mathbf{M}^T$, where $\mathbf{M}$ is orthogonal and $\mathbf{\Sigma}$ is a diagonal matrix consisting of the eigenvalues of $\mathbf{C}$. Because $\operatorname{trace}(\mathbf{C}) = 1$, the sum of the eigenvalues on the diagonal of $\mathbf{\Sigma}$ is 1. Moreover, all eigenvalues are positive. If all eigenvalues are identical, we estimate a uniform distribution with arbitrary $\underline{\mu}$ and $\kappa = 0$. Otherwise, we proceed as follows.

If we assume that $\mathbf{C}$ stems from a Watson distribution, we find that $d-1$ eigenvalues are identical and one eigenvalue σ is different. We then obtain $\kappa = D_d^{-1}(\sigma)$ and get $\underline{\mu}$ as the corresponding eigenvector.

For an arbitrary $\mathbf{C}$, it is not possible to match the covariance exactly because there may be more distinct eigenvalues. In this case, we consider the smallest and the largest eigenvalues σ_1 and σ_2, respectively. We then obtain two candidates for κ as $\kappa_1 = D_d^{-1}(\sigma_1)$ and $\kappa_2 = D_d^{-1}(\sigma_2)$. To pick one of the candidates, we consider

$$\arg\max_{j \in \{1,2\}} \left(\kappa_j \sigma_j - \log M(1/2, d/2, \kappa_j) \right) .$$

This condition is motivated by the fact that it yields the same result as before if $\mathbf{C}$ stems from a Watson distribution and that it also matches the result that would be obtained by a maximum likelihood estimator. The parameter $\underline{\mu}$ is obtained as the eigenvector corresponding to the chosen κ.

The proposed moment-based estimator is closely-related to the moment-based estimator for the Bingham distribution [35, Sec. 5.2]. However, in the case of the Bingham distribution, it is always possible to match the second moment because the Bingham distribution has more degrees of freedom.

References

[1] S. Kullback, *Information Theory and Statistics*. Dover, 1978.

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: MIT press, 2012.

[3] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[4] R. Herbrich, "Minimising the Kullback–Leibler Divergence," Microsoft, Tech. Rep., Aug. 2005. [Online]. Available: http://research.microsoft.com/pubs/74555/KL.pdf

[5] A. Chiuso and G. Picci, "Visual Tracking of Points as Estimation on the Unit Sphere," in *The Confluence of Vision and Control*. Springer, 1998, vol. 237, pp. 90–105.

[6] G. Kurz and U. D. Hanebeck, "Trigonometric Moment Matching and Minimization of the Kullback–Leibler Divergence," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 1, pp. 3480–3484, Oct. 2015.

[7] C. Bingham, "An Antipodally Symmetric Distribution on the Sphere," *The Annals of Statistics*, vol. 2, no. 6, pp. 1201–1225, Nov. 1974.

[8] G. S. Watson, "Equatorial Distributions on a Sphere," *Biometrika*, vol. 52, no. 1–2, pp. 193–201, 1965.

[9] E. Dimroth, "Untersuchungen zum Mechanismus von Blastesis und Syntexis in Phylliten und Hornfelsen des südwestlichen Fichtelgebirges – I. Die statistische Auswertung einfacher Gürteldiagramme," *Mineralogy and Petrology*, vol. 8, no. 2, pp. 248–274, 1962.

[10] I. Markovic, F. Chaumette, and I. Petrovic, "Moving Object Detection, Tracking and Following Using an Omnidirectional Camera on a Mobile Robot," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA 2014)*, Hong-Kong, Jun. 2014.

[11] I. Markovic, M. Bukal, J. Cesic, and I. Petrovic, "Direction-only Tracking of Moving Objects on the Unit Sphere via Probabilistic Data Association," in *Proceedings of the 17th International Conference on Information Fusion (Fusion 2014)*, Salamanca, Spain, Jul. 2014.

[12] Y.-H. Chen, D. Wei, G. Newstadt, M. DeGraef, J. Simmons, and A. Hero, "Parameter Estimation in Spherical Symmetry Groups," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1152–1155, 2015.

[13] J. Traa and P. Smaragdis, "Multiple Speaker Tracking With the Factorial von Mises–Fisher Filter," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2014.

[14] H. Tang, S. Chu, and T. Huang, "Generative Model-based Speaker Clustering via Mixture of von Mises–Fisher Distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, 2009, pp. 4101–4104.

[15] G. Kurz, I. Gilitschenski, S. Julier, and U. D. Hanebeck, "Recursive Bingham Filter for Directional Estimation Involving 180 Degree Symmetry," *Journal of Advances in Information Fusion*, vol. 9, no. 2, pp. 90–105, Dec. 2014.

[16] J. Glover and L. P. Kaelbling, "Tracking the Spin on a Ping Pong Ball with the Quaternion Bingham Filter," in *Proceedings of the 2014 IEEE Conference on Robotics and Automation (ICRA 2014)*, Hong Kong, China, 2014.

[17] J. T. Kent and T. Hamelryck, "Using the Fisher-Bingham Distribution in Stochastic Models for Protein Structure," *Quantitative Biology, Shape Analysis, and Wavelets*, vol. 24, pp. 57–60, 2005.

[18] A. S. Bijral, M. Breitenbach, and G. Z. Grudic, "Mixture of Watson Distributions: A Generative Model for Hyperspherical Embeddings," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 35–42.

[19] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using von Mises–Fisher Distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[20] P. Leong and S. Carlile, "Methods for Spherical Data Analysis and Visualization," *Journal of Neuroscience Methods*, vol. 80, no. 2, pp. 191–200, 1998.

[21] G. Kurz, I. Gilitschenski, and U. D. Hanebeck, "Unscented von Mises-Fisher Filtering," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 463–467, Apr. 2016.

[22] J. E. Darling and K. J. DeMars, "Minimization of the Kullback–Leibler Divergence for Nonlinear Estimation," in *Proceedings of the Astrodynamics Specialist Conference*, Vail, Colorado, USA, Aug. 2015.

[23] B. Chen, Y. Zhu, J. Hu, and Z. Sun, "Adaptive Filtering Under Minimum Information Divergence Criterion," *International Journal of Control, Automation and Systems*, vol. 7, no. 2, pp. 157–164, 2009.

[24] R. Fisher, "Dispersion on a Sphere," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 217, no. 1130, pp. 295–305, 1953.

[25] R. von Mises, "Über die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen," *Physikalische Zeitschrift*, vol. XIX, pp. 490–500, 1918.

[26] G. S. Watson, "Large sample theory of the Langevin distribution," *Journal of Statistical Planning and Inference*, vol. 8, no. 3, pp. 245–256, 1983.

[27] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed. New York: Dover, 1972.

[28] A. Pewsey, M. Neuhäuser, and G. D. Ruxton, *Circular Statistics in R*, 1st ed. Oxford, UK: Oxford University Press, 2013.

[29] S. Sra, "A Short Note on Parameter Approximation for von Mises–Fisher Distributions: And a Fast Implementation of Is(x)," *Computational Statistics*, vol. 27, no. 1, pp. 177–190, 2012.

[30] K. V. Mardia and P. E. Jupp, *Directional Statistics*, 1st ed. Baffins Lane, Chichester, West Sussex, England: Wiley, 1999.

[31] D. E. Amos, "Computation of Modified Bessel Functions and Their Ratios," *Mathematics of Computation*, vol. 28, no. 125, pp. 239–251, 1974.

[32] S. Sra and D. Karp, "The Multivariate Watson Distribution: Maximum-Likelihood Estimation and other Aspects," *Journal of Multivariate Analysis*, vol. 114, pp. 256–269, 2013.

[33] J. Glover and L. P. Kaelbling, "Tracking 3-D Rotations with the Quaternion Bingham Filter," MIT, Tech. Rep., Mar. 2013.

[34] G. Kurz and U. D. Hanebeck, "Stochastic Sampling of the Hyperspherical von Mises–Fisher Distribution Without Rejection Methods," in *Proceedings of the IEEE ISIF Workshop on Sensor Data Fusion: Trends, Solutions, Applications (SDF 2015)*, Bonn, Germany, Oct. 2015.

[35] ——, "Dynamic Surface Reconstruction by Recursive Fusion of Depth and Position Measurements," *Journal of Advances in Information Fusion*, vol. 9, no. 1, pp. 13–26, Jun. 2014.