# Directional Statistics with the Spherical Normal Distribution

Søren Hauberg
*Department of Mathematics and Computer Science*
*Technical University of Denmark*
Kgs. Lyngby, Denmark
sohau@dtu.dk

***Abstract*—A well-known problem in directional statistics — the study of data distributed on the unit sphere — is that current models disregard the curvature of the underlying sample space. This ensures computationally efficiency, but can influence results. To investigate this, we develop efficient inference techniques for data distributed by the curvature-aware *spherical normal distribution*. We derive closed-form expressions for the normalization constant when the distribution is isotropic, and a fast and accurate approximation for the anisotropic case on the two-sphere. We further develop approximate posterior inference techniques for the mean and concentration of the distribution, and propose a fast sampling algorithm for simulating the distribution. Combined, this provides the tools needed for practical inference on the unit sphere in a manner that respects the curvature of the underlying sample space.**

## I. Introduction

*Directional statistics* considers data distributed on the unit sphere, i.e. $\|\mathbf{x}_n\| = 1$. The corner-stone model is the *von Mises-Fisher distribution* [19] which is derived by restricting an isotropic normal distribution to lie only on the unit sphere:

$$\mathsf{vMF}(\mathbf{x} \mid \boldsymbol{\mu}, \kappa) \propto \exp\left(-\frac{\kappa}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right) \propto \exp\left(\kappa\ \mathbf{x}^\intercal\boldsymbol{\mu}\right), \\ \|\mathbf{x}\| = \|\boldsymbol{\mu}\| = 1, \quad \kappa > 0. \tag{1}$$

Evidently, the von Mises-Fisher distribution is constructed from the Euclidean distance measure $\|\mathbf{x} - \boldsymbol{\mu}\|$, i.e. the length of the connecting straight line. A seemingly more natural choice of distance measure is the arc-length of the shortest connecting curve on the sphere, which amounts to the angle between two points, i.e. $\mathrm{dist}(\mathbf{x}, \boldsymbol{\mu}) = \arccos(\mathbf{x}^\intercal\boldsymbol{\mu})$. From this measure, it is natural to consider the distribution

$$\mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \lambda) \propto \exp\left(-\frac{\lambda}{2}\arccos^2(\mathbf{x}^\intercal\boldsymbol{\mu})\right), \\ \|\mathbf{x}\| = \|\boldsymbol{\mu}\| = 1, \quad \lambda > 0, \tag{2}$$

with concentration parameter $\lambda$. This is an instance of the *Riemannian normal distribution* [20] which is naturally deemed the *spherical normal distribution*. This is the topic of interest in the present manuscript.

At this point, the reader may wonder if the choice of distance measure is purely academic or if it influences practical inference. We therefore consider a simple numerical experiment, where we evaluate the likelihood of the mean $\boldsymbol{\mu}$ of a von Mises-Fisher and a spherical normal distribution. We consider one observation on the north pole, and another observations which we move along a great circle from the north to the south pole. When the two observations are on opposite poles the mean $\boldsymbol{\mu}$ cannot be expected to be unique; rather intuition dictate that any point along the equator (the set of points that are equidistant to the poles) is a suitable candidate mean. Figure 1 shows the experiment along with the numerically evaluated likelihood of $\boldsymbol{\mu}$. We observe that under the von Mises-Fisher distribution, this likelihood is *isotropic* with an increasing variance as the observations move further apart. When the observations are on opposite poles, the likelihood becomes uniform over the entire sphere, implying that any choice of mean is as good as another. This is a rather surprising result, that align poorly with geometric intuition. Under the spherical normal distribution, the likelihood is seen to be *anisotropic*, where the variance increase more orthogonally to the followed great circle, than it does along the great circle. Finally, when the observation reaches the south pole, the likelihood concentrates in a "belt"

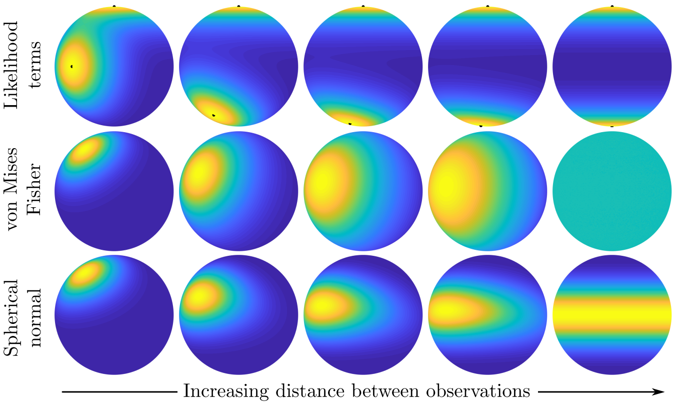

Fig. 1. Likelihood of the spherical mean under the von Mises-Fisher and spherical normal models. *Top row:* Two observations are moved further and further apart until they are on opposite poles (coloring show the sum of the likelihood terms for the data). *Middle row:* The likelihood of the mean under a von Mises-Fisher distribution. As the data moves further apart, the variance of the likelihood grows until it degenerates into the uniform distribution. *Bottom row:* The likelihood of the mean under a spherical normal distribution. As the observations moves apart, the likelihood becomes increasingly anisotropic until it stretches along the equator. This is the most natural result.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360). SH was supported by a research grant (15334) from VILLUM FONDEN.

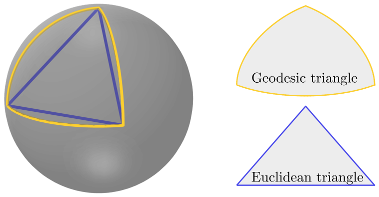

Fig. 2. Flat versus curved metrics. Straight lines (purple) correspond to Euclidean distances used by the von Mises-Fisher distribution. This *flat metric* is in contrast to the spherical arc-length distance (yellow geodesic curves). Under this distance measure, geodesic triangles are "fatter" than corresponding Euclidean triangles. This distortion, which is the key characteristics of the sphere, is disregarded by the von Mises-Fisher.

along the equator, implying that any mean along the equator is a good candidate. This likelihood coincide with the geometric intuition.

The core issue is that the Euclidean distance measure applied by the von Mises-Fisher distribution is a *flat metric* [8] as illustrated in Fig. 2. This imply that the von Mises-Fisher distribution is incapable of reflecting the curvature of the sphere. As curvature is one of the defining mathematical properties of the spherical sample space, it is a serious limitation of the von Mises-Fisher model that this is disregarded as it may give misleading results. For practical inference, there is, however, no viable alternative that respect the curvature of the spherical sample space. We fill this gap in the literature and contribute

- closed-form expressions for the normalization constant of the isotropic spherical normal distribution (Sec. III-A) and an efficient and accurate approximation for anisotropic case on the two-sphere;
- maximum likelihood (Sec. IV) and approximate Bayesian inference techniques (Sec. V) for the spherical normal;
- an efficient sampling algorithm for the spherical normal (Sec. VI).

As examples of these techniques, we provide a new clustering model on the sphere (Sec. IV-C) and a new algorithm for Kalman filtration on the sphere (Sec. V-B). Relevant source code will be published alongside this manuscript.

## II. Background and Related Work

*a) Spherical geometry.:* We start the exposition with a brief review of the geometry of the unit sphere. This is a compact Riemannian manifold with constant unit curvature. It has surface area $A_{D-1} = \frac{2\pi^{D/2}}{\Gamma({}^D/_2)}$, where $\Gamma$ is the usual Gamma function. It is often convenient to linearize the sphere around a base point $\boldsymbol{\mu}$, such that points on the sphere are represented in a local Euclidean tangent space $\mathcal{T}_{\boldsymbol{\mu}}$ (Fig. 3). The convenience stems from the linearity of tangent space, but this comes at the cost that $\mathcal{T}_{\boldsymbol{\mu}}$ gives a distorted view of the curved sphere. Since the maximal distance from $\boldsymbol{\mu}$ to any point is $\pi$, we are generally only concerned with tangent vectors $\mathbf{v} \in \mathcal{T}_{\boldsymbol{\mu}}$ where $\|\mathbf{v}\| \leq \pi$. A point $\mathbf{x} \in \mathcal{S}^{D-1}$ can be mapped to $\mathcal{T}_{\boldsymbol{\mu}}$ via the

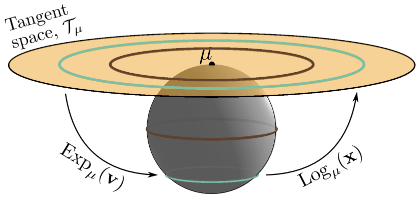
Tangent space, $\mathcal{T}_\mu$
$\mu$
$\mathrm{Exp}_\mu(\mathbf{v})$
$\mathrm{Log}_\mu(\mathbf{x})$

Fig. 3. The tangent space $\mathcal{T}_{\boldsymbol{\mu}}$ of the sphere and related operators.

*logarithm map*,

$$\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}\,(\mathbf{x}^{\intercal}\boldsymbol{\mu}))\frac{\theta}{\sin(\theta)}, \qquad \theta = \arccos(\mathbf{x}^{\intercal}\boldsymbol{\mu}), \tag{3}$$

with the convention ${}^0/_{\sin(0)} = 1$. The inverse mapping, known as the *exponential map*, moves a tangent vector $\mathbf{v}$ back to the sphere,

$$\mathrm{Exp}_{\boldsymbol{\mu}}(\mathbf{v}) = \boldsymbol{\mu}\cos(\|\mathbf{v}\|) + {}^{\sin(\|\mathbf{v}\|)}/_{\|\mathbf{v}\|}\mathbf{v}. \tag{4}$$

A tangent vector $\mathbf{v} \in \mathcal{T}_{\boldsymbol{\mu}}$ can be moved to another tangent space $\mathcal{T}_{\boldsymbol{\mu}'}$ by a *parallel transport*. This amounts to applying a rotation $\mathbf{R}$ that move $\boldsymbol{\mu}$ to $\boldsymbol{\mu}'$ [9].

*b) Distributions on the sphere.:* Since the unit sphere is a compact space, we can define a uniform distribution with density

$$\mathsf{Uniform}(\mathbf{x}) = A_{D-1}^{-1} = \frac{\Gamma({}^D/_2)}{2\pi^{D/2}}. \tag{5}$$

The corner-stone distribution on the unit sphere is the already discussed von Mises-Fisher distribution [19], which is derived by restricting the isotropic normal distribution to the unit sphere. Consequently it is defined with respect to the Euclidean distance measure. The mean parameter of the von Mises-Fisher distribution can be estimated with maximum likelihood in closed-form, but the concentration parameter $\kappa$ must be estimated using numerical optimization [19]. Bayesian analysis is simplified since the distribution is conjugate with itself for the mean, and with the Gamma distribution for the concentration.

The von Mises-Fisher distribution can be extended to be anisotropic by restricting an anisotropic normal distribution to the unit sphere [16], giving the *Fisher-Bingham* (or *Kent*) distribution,

$$p(\mathbf{x}) = \frac{1}{\mathcal{C}}\exp\left(\kappa\boldsymbol{\gamma}_1^{\intercal}\mathbf{x} + \beta[(\boldsymbol{\gamma}_2^{\intercal}\mathbf{x})^2 - (\boldsymbol{\gamma}_3^{\intercal}\mathbf{x})^2]\right), \tag{6}$$

where the normalization constant $\mathcal{C}$ can be evaluated in closed-form on $\mathcal{S}^2$. Like the von Mises-Fisher distribution, this use the Euclidean distance function, and therefore does not respect the curvature of the sphere.

One approach to constructing an anisotropic curvature-aware distribution is to assume that $\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})$ follow a zero-mean anisotropic normal distribution [28]. This *tangential normal distribution* is simple, and since it is defined over the Euclidean tangent space, standard Bayesian analysis hold for its precision (i.e. it is conjugate with the Wishart distribution). However, the distribution disregard the distortion of the tangent space: following the *change of variable theorem* we see that the tangential normal over $\mathcal{S}^2$ has density $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left|\det(\partial_{\mathbf{x}} \mathrm{Log}_{\boldsymbol{\mu}})\right| = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, {}^{\theta}/_{\sin(\theta)}$, where $\theta = \arccos(\mathbf{x}^\intercal \boldsymbol{\mu})$. Surprisingly, this distribution is thus bimodal with modes at $\pm\boldsymbol{\mu}$. This rather unintuitive property indicate that one should be careful when applying this model.

An approach related to ours is that of Purkayastha [21] who study $p(\mathbf{x}) \propto \exp(-\kappa \arccos(\mathbf{x}^\intercal \boldsymbol{\mu}))$ and show that the maximum likelihood estimate of $\boldsymbol{\mu}$ is the intrinsic median on the sphere.

*c) Manifold statistics.:* The spherical normal distribution, which we consider in this manuscript, is an instance of the general *Riemannian normal distribution*. While this distribution has seen significant theoretical studies [20, 29], practical tools for inference are lacking. Generally, its normalization constant depend on both mean and concentration parameters and is unattainable in closed-form. Even evaluating the gradient of the log-likelihood with respect to the mean, thus, require expensive Monte Carlo schemes [5, 31]. Similar concerns hold for the concentration parameters. A key contribution of this paper is that we provide tools for the spherical setting that drastically simplify and speed up parameter estimation. We further provide tools for approximate Bayesian inference on Riemannian manifolds — a topic that has yet to be addressed in the literature.

*d) Applications of directional statistics.:* The use of directional distributions span many branches of science, ranging from modeling *wind directions* [19], *protein shapes* [10], *gene expressions* [6] to *fMRI time series* [22]. In these applications, the directional model appear due to normalization of the data. *Histograms*, such as *word counts*, can also be efficiently modeled as directional data [6]; in fact, the arc-length distance on the sphere correspond to the distance between discrete distributions under the Fisher-Rao metric [4], which is another indication that we need curvature-aware models. Other recent applications include *clustering surface normals* to *aide robot navigation* [27, 28], and *speaker direction tracking* [30]. At the foundational level, directional statistics also appear in probabilistic estimates of the *principal components* of Euclidean data [14, 26].

## III. Basic properties of the spherical normal

The spherical normal distribution as presented in Eq. 2 is isotropic. Noting that $\arccos^2(\mathbf{x}^\intercal \boldsymbol{\mu}) = \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})^\intercal \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})$ is merely the squared Euclidean norm measured in the tangent space $\mathcal{T}_{\boldsymbol{\mu}}$ allows us to generalize Eq. 2 to be anisotropic:

$$\mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp\left(-\frac{1}{2} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})^\intercal \boldsymbol{\Lambda} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})\right), \tag{7}$$

where the *concentration parameter* $\boldsymbol{\Lambda}$ is a symmetric positive semidefinite matrix. This is defined in $\mathcal{T}_{\boldsymbol{\mu}}$ and consequently has an eigenvector $\boldsymbol{\mu}$ with eigenvalue zero.

When the concentration goes towards zero, it is easy to see that the spherical normal tends towards the uniform distribution, i.e. when $\boldsymbol{\Lambda} = \lambda \boldsymbol{I}$ we have

$$\lim_{\lambda \to 0^+} \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \lambda \boldsymbol{I}) = \mathsf{Uniform}(\mathbf{x}). \tag{8}$$

When $\lambda$ goes towards infinity, the spherical normal becomes a delta function. Similar properties hold for the von Mises-Fisher distribution [19].

It is worth noting that this general spherical normal distribution is the spherical distribution with maximal entropy for a given mean and covariance [20]. Note that this is with respect to the probability measure induced by the spherical metric. Similar statements hold for the von Mises-Fisher distribution [19], but with respect to the measure induced by the Euclidean metric.

The spherical normal, thus, share many desirable properties with the von Mises-Fisher distribution, while still being curvature-aware.

### A. Normalizing the spherical normal

Equation 2 only specifies the spherical normal distribution up to a constant. In the online supplements[1] we show the following result:

**Proposition 1** (Isotropic normalization)**.** *The normalization constant of the isotropic spherical normal distribution* (2) *over* $\mathcal{S}^{D-1}$ *is given by Eq. 9a when* $D$ *is even and by Eq. 9b when* $D$ *is odd (see Fig. 4). Here* erf *is the imaginary error function [1], and* $\mathrm{Re}[\cdot]$ *and* $\mathrm{Im}[\cdot]$ *takes the real and imaginary parts of a complex number, respectively.*

**Remark 1.** *While the complex unit appear in Eq. 9 the entire expression evaluates to a real number.*

The anisotropic spherical normal distribution (7) does, unfortunately, not appear to have a closed-form expression for its normalization constant, and we must resort to approximations. Here we consider the anisotropic spherical normal over the ever-present $\mathcal{S}^2$ and provide a simple deterministic approximation that, in our experience, provides a relative error of approximately $10^{-4}$. This is more than sufficient for practical tasks. To derive this, we first express the distribution in the tangent space of the mean,

$$\mathcal{Z}_2(\boldsymbol{\Lambda}) = \int_{\mathcal{S}^2} \exp\left(-\frac{1}{2} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})^\intercal \boldsymbol{\Lambda} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \tag{10}$$

$$= \int_{\|\mathbf{v}\| \leq \pi} \exp\left(-\frac{1}{2} \mathbf{v}^\intercal \boldsymbol{\Lambda} \mathbf{v}\right) \det(\mathbf{J}) \mathrm{d}\mathbf{v}, \tag{11}$$

where $\mathbf{v} = \mathrm{Log}_{\mu}(\mathbf{x})$ and $\det(\mathbf{J}) = {}^{\sin(\|\mathbf{v}\|)}/_{\|\mathbf{v}\|}$ denotes the Jacobian of $\mathrm{Exp}_{\mu}(\mathbf{x})$. This expression is independent of the choice of orthogonal basis of the tangent space, so we need only consider the case where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, 0)$. Writing Eq. 11 in polar coordinates $(r, \theta)$ then shows

$$\begin{aligned}\mathcal{Z}_2(\boldsymbol{\Lambda}) = \int_0^{2\pi} \int_0^{\pi} \exp\Big\{ -\frac{r^2}{2} (\lambda_1 \cos^2(\theta) \\ + \lambda_2 \sin^2(\theta)) \Big\} \sin(r) \mathrm{d}r \mathrm{d}\theta\end{aligned} \tag{12}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \mathcal{Z}_1\left((\lambda_1 - \lambda_2)\cos^2(\theta) + \lambda_2\right) \mathrm{d}\theta. \tag{13}$$

[1] http://www2.compute.dtu.dk/~sohau/papers/fusion2018/hauberg_fusion2018_supplements.pdf

$$\mathcal{Z}_1^{(\text{even})}(\lambda) = \frac{A_{D-2}}{2^{D-2}}\binom{D-2}{{}^{D}\!/_{2}-1}\sqrt{\frac{\pi}{2\lambda}}\text{erf}\left(\pi\sqrt{\frac{\lambda}{2}}\right) + \frac{A_{D-2}}{2}\sqrt{\frac{2\pi}{\lambda}}\frac{(-1)^{{}^{D}\!/_{2}-1}}{2^{D-3}}$$

$$\cdot \sum_{k=0}^{{}^{D}\!/_{2}-2}(-1)^k\binom{D-2}{k}\exp\left(-\frac{(D-2-2k)^2}{2\lambda}\right)\text{Re}\left[\text{erf}\left(\frac{\pi\lambda-(D-2-2k)i}{\sqrt{2\lambda}}\right)\right] \tag{9a}$$

$$\mathcal{Z}_1^{(\text{odd})}(\lambda) = A_{D-2}\frac{(-1)^{(D\,-\,3)/2}}{2^{D-3}}\sqrt{\frac{\pi}{2\lambda}}\sum_{k=0}^{(D\,-\,3)/2}(-1)^k\binom{D-2}{k}\exp\left(-\frac{(D-2-2k)^2}{2\lambda}\right)$$

$$\cdot\left\{\text{Im}\left[\text{erf}\left(\frac{(D-2-2k)i}{\sqrt{2\lambda}}\right)\right]+\text{Im}\left[\text{erf}\left(\frac{\pi\lambda-(D-2-2k)i}{\sqrt{2\lambda}}\right)\right]\right\} \tag{9b}$$

Fig. 4. Normalization constants for the isotropic spherical normal distribution, when $D$ is even and odd. Here erf is the imaginary error function, while Re[$\cdot$] and Im[$\cdot$] takes the real and imaginary parts of a complex number, respectively.

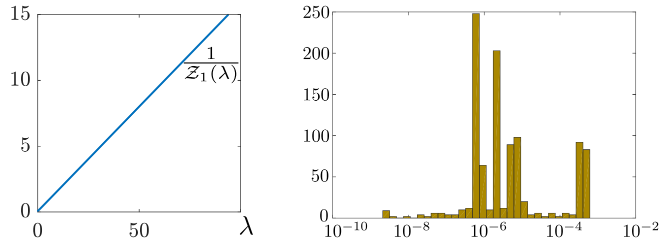

Fig. 5. *Left:* The inverse of $\mathcal{Z}_1(\lambda)$. *Right:* Relative errors when computing $\mathcal{Z}_2(\mathbf{\Lambda})$.

This does not appear to have a closed-form expression, but does lend itself to a good approximation. Noting that ${}^{1}\!/_{\mathcal{Z}_1(\lambda)}$ is almost a straight line (see left panel of Fig. 5), we approximate

$$\mathcal{Z}_2(\mathbf{\Lambda}) \approx \int_0^{2\pi}\frac{1}{a\left((\lambda_1-\lambda_2)\cos^2(\theta)+\lambda_2\right)+b}\mathrm{d}\theta \tag{14}$$

$$= \frac{1}{\sqrt{\det\left(a\mathbf{\Lambda}+b\mathbf{I}\right)}}, \tag{15}$$

where $a$ and $b$ are estimated by fitting a straight line to ${}^{1}\!/_{\mathcal{Z}_1(\lambda)}$. In Sec. IV-C we consider an EM algorithm for mixtures of spherical normal distributions. Here we track the relative approximation error of the normalization constant (using expensive numerical integration as ground truth) through the entire run of the EM algorithm. The right panel of Fig. 5 shows a histogram of these errors. From this data, we observe that the relative error of our proposed approximation is between $2.1 \cdot 10^{-9}$ and $6.1 \cdot 10^{-4}$ with an average of $7.9 \cdot 10^{-5}$. This is plenty accurate for practical inference tasks.

### B. Rotations and convolutions

In the spherical domain we cannot perform addition and multiplication, but similar operations are available. Here we state without proof that spherical normals are closed under rotation

$$\mathbf{x} \sim \mathsf{SN}(\boldsymbol{\mu}, \mathbf{\Lambda}) \quad \Rightarrow \quad \mathbf{Rx} \sim \mathsf{SN}(\boldsymbol{\mu}, \mathbf{R\Lambda R}^{\intercal}), \tag{16}$$

where $\mathbf{R}$ is a rotation matrix. Two independent spherical normally distributed variables can (informally) be "added" by convolving their distributions. When these variables have the same mean, then

$$\left.\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Lambda}_{\mathbf{x}}) \\ p_{\mathbf{y}}(\mathbf{y}) &= \mathsf{SN}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{\Lambda}_{\mathbf{y}}) \end{aligned}\right\} \Rightarrow \tag{17}$$

$$(p_{\mathbf{x}} \star p_{\mathbf{y}})(\mathbf{z}) = \mathsf{SN}\left(\mathbf{z} \mid \boldsymbol{\mu},\ \mathbf{\Lambda}_{\mathbf{x}} + \mathbf{\Lambda}_{\mathbf{y}}\right),$$

where $\star$ denotes convolution. In Sec. V-B we use both results in a spherical Kalman filter.

### C. A note on covariances

The dispersion of the spherical normal is parametrized by a concentration matrix $\mathbf{\Lambda}$. From this the variance and covariance of the distribution is defined as [20]

$$\mathsf{Var}[\mathbf{x}] = \int_{\mathcal{S}^{D-1}} \arccos^2(\mathbf{x}^{\intercal}\boldsymbol{\mu})\ \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Lambda})\mathrm{d}\mathbf{x} \tag{18}$$

$$\mathsf{Cov}[\mathbf{x}] = \int_{\mathcal{S}^{D-1}} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})^{\intercal}\ \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Lambda})\mathrm{d}\mathbf{x}.$$

The empirical counterparts of these expressions are found by replacing integrals with averages as usual. Generally we can evaluate these expressions numerically by sampling (Sec. VI), though we note that in the case of an isotropic spherical normal over $\mathcal{S}^2$, the variance can be expressed in closed-form (see supplements[1]). The expression is, however, somewhat more convoluted than in the Euclidean setting where the variance is the inverse concentration (precision). Note that the variance is bounded by $\pi^2$ since the maximal spherical distance is $\pi$.

## IV. Maximum likelihood estimation

We now consider maximum likelihood estimation for the spherical normal. As usual, we write the log-likelihood of data $\mathbf{x}_{1:N}$

$$f(\boldsymbol{\mu}, \mathbf{\Lambda}) = \log\left(\prod_{n=1}^{N}\mathsf{SN}(\mathbf{x}_n \mid \boldsymbol{\mu}, \mathbf{\Lambda})\right) \tag{19}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^{\intercal}\mathbf{\Lambda}\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n) - N\log(\mathcal{Z}_2(\mathbf{\Lambda})).$$

**Algorithm 1** Offline maximum likelihood: $\boldsymbol{\mu}$

```
Initialize $\boldsymbol{\mu}$ as a random data point.
repeat
    $\nabla_{\boldsymbol{\mu}} \leftarrow -\frac{1}{N}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)$.
    $\boldsymbol{\mu} \leftarrow \mathrm{Exp}_{\boldsymbol{\mu}}\left(-\frac{1}{2}\nabla_{\boldsymbol{\mu}}\right)$.
until $\|\nabla_{\boldsymbol{\mu}}\| < 10^{-5}$.
```

**Algorithm 2** Online maximum likelihood: $\boldsymbol{\mu}$

```
Initialize $\boldsymbol{\mu} \leftarrow \mathbf{x}_1$.
for $n = 2$ to $\ldots$ do
    $\boldsymbol{\mu} \leftarrow \mathrm{Exp}_{\boldsymbol{\mu}}\left(\frac{1}{n}\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)\right)$.
end for
```

### A. The mean

Since the normalization constant does not depend on $\boldsymbol{\mu}$, it is easy to see that the derivative of the log-likelihood with respect to the mean is

$$\frac{\partial f}{\partial \boldsymbol{\mu}} = -\frac{1}{N}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^{\intercal}\boldsymbol{\Lambda}. \tag{20}$$

The mean can then be found by standard Riemannian gradient descent [2], but it is easier to consider the steepest descent [5] given by $\nabla_{\boldsymbol{\mu}} = -\frac{1}{N}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)$. The optimization then amounts to mapping the data to the tangent at the current $\boldsymbol{\mu}$, computing its Euclidean average, and mapping it back to the sphere. This is summarized in Algorithm 1.

Alternatively, the mean can be computed by repeated geodesic interpolation [23], which can be done in an online fashion. From two points, the mean is estimated as the midpoint of the connecting geodesic; when given a third point, the mean is updated by going $1/3$ along the geodesic between the two-point mean and the third point, and so forth. This is summarized in Algorithm 2. Salehian et al. [23] show that this simple estimator converges to the true mean. This algorithm can also be seen as stochastic gradient descent with a particularly efficient step-size selection.

### B. The concentration matrix

Using our approximation of the normalization constant (15), the gradient of the log-likelihood with respect to $\boldsymbol{\Lambda}$ is (expressed in a two-dimensional basis of $\mathcal{T}_{\boldsymbol{\mu}}$)

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{\Lambda}} = -\frac{1}{2}\Bigg\{ &\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^{\intercal} \\ &- aN\,\left(a\,\boldsymbol{\Lambda} + b\,\mathbf{I}\right)^{-1}\Bigg\}.\end{aligned} \tag{21}$$

An optimum can be found using standard Riemannian optimization [2] over the symmetric positive definite cone. Initializing with the inverse covariance of $\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})$ usually ensures convergence in $3-4$ iterations. This initialization coincides with the standard "least squares" estimator of the inverse covariance on general Riemannian manifolds [5]. It is therefore not surprising that only few iterations are required, and when computational speed is of the utmost importance, this estimator is a good approximation. Pennec [20] provides an alternative approximate estimator for small concentration matrices that compensates for the curvature of the sample space.

### C. Example: EM algorithm for mixture models

It is straight-forward to extend the presented maximum likelihood estimators to mixture models, i.e. $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$, where the responsibilities $\pi_k$ satisfy $\sum_k \pi_k = 1, \pi_k \geq 0$. The derivation follows the well-known Euclidean analysis [7] and will not be repeated here.

As an example, we consider clustering of surface normals extracted from depth images [11, 25]. This is, e.g., useful for robot navigation [27, 28]. In practice, such data is highly noisy due to limitations of the depth-sensing camera, so we extend the mixture model to include a component with uniform distribution, i.e. $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathsf{SN}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \pi_{K+1} \cdot \mathsf{Uniform}(\mathbf{x})$. The left panel of Fig. 6 shows example data, the contours of the estimated spherical normals, and the corresponding regions on the sphere where each component is dominant. The center panels show the corresponding image segmentation. This seems to correspond well to the scene geometry. It is interesting to note that the uniform component mostly captures edge and shadow areas where surface normals are unstable. In the figure, we consider 5 spherical normal components.

Next we sample 50 random depth images from the *NYU Depth Dataset* [25], which depicts indoor environments. In such scenes, a few surface normals tend to dominate as most furniture has similar normals as the walls, floors and ceilings. We perform clustering on these 50 images using both a mixture of spherical normal distributions and a mixture of von Mises-Fisher distributions. For each model, we select the optimal number of components using the *Akaike Information Criteria (AIC)* [3] measured on held-out data. The right-most panel of Fig. 6 show a histogram of the number of selected components for both mixture models. We see that significantly fewer components are needed when using a mixture of spherical normals than when using a mixture of von Mises-Fisher distributions. This indicate that the von Mises-Fisher model tend to over-segment the depth images.

## V. Approximate Bayesian inference

Thus far, we have considered maximum likelihood estimation for the spherical normal, however, some may prefer a Bayesian machinery. Here we consider the mean and concentration separately.

### A. Unknown mean, known concentration

Setting a spherical normal prior over the mean $p(\boldsymbol{\mu}) = \mathsf{SN}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ while assuming a spherical normal likelihood $p(\mathbf{x}_{1:N} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_n \mathsf{SN}(\mathbf{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Lambda})$ quickly reveals that the resulting posterior is *not* a spherical normal. This suggests

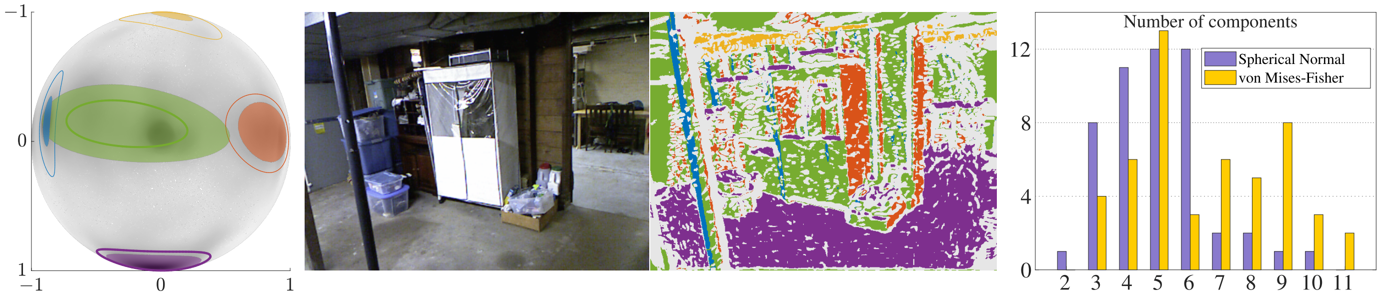

Fig. 6. Clustering of surface normals; see text for details.

that approximations are in order. First, we note that the log-posterior,

$$g(\boldsymbol{\mu}) = \log(p(\boldsymbol{\mu} \mid \mathbf{x}_{1:N}, \boldsymbol{\Lambda}, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)) \tag{22}$$

$$\begin{aligned} &= -\frac{1}{2}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^\intercal \boldsymbol{\Lambda} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n) \\ &\quad - \frac{1}{2}\mathrm{Log}_{\boldsymbol{\mu}_0}(\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}_0 \mathrm{Log}_{\boldsymbol{\mu}_0}(\boldsymbol{\mu}) + \text{const}, \end{aligned} \tag{23}$$

is simultaneously expressed in tangent spaces at both $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$, which complicates the analysis. Due to the symmetry of the spherical normal, we can, however, rewrite Eq. 23 as

$$\begin{aligned} g(\boldsymbol{\mu}) = &-\frac{1}{2}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^\intercal \boldsymbol{\Lambda} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n) \\ &- \frac{1}{2}\mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0)^\intercal \mathbf{R}\boldsymbol{\Lambda}_0\mathbf{R}^\intercal \mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0) + \text{const}, \end{aligned} \tag{24}$$

where $\mathbf{R}$ is the parallel transport from $\mathcal{T}_{\boldsymbol{\mu}_0}$ to $\mathcal{T}_{\boldsymbol{\mu}}$. The maximum a posteriori (MAP) estimate of the mean can then be found with Riemannian gradient descent, where the gradient is

$$\frac{\partial g}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^\intercal \boldsymbol{\Lambda} - \frac{1}{2}\mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0)^\intercal \mathbf{R}\boldsymbol{\Lambda}_0\mathbf{R}^\intercal. \tag{25}$$

The optimal $\boldsymbol{\mu}_N$ can be found akin to Algorithm 1. This is, however, only a point estimate and does not approximate the full posterior. Here we consider a Laplace approximation [7] as these generalize easily to the spherical setting. Since the metric in $\mathcal{T}_{\boldsymbol{\mu}}$ reduces to the identity around the origin, this amounts to approximating the concentration of the posterior $\boldsymbol{\Lambda}_N$ with the Riemannian Hessian of $g(\boldsymbol{\mu})$ at $\boldsymbol{\mu}_N$ [2].

Figure 7A shows an example with a single observation with a spherical normal likelihood, as well as a spherical normal prior over the unknown mean. The intensity of the sphere is a numerical evaluation of the true posterior, while the orange curve outlines the Laplace approximation of the mean posterior. This appears to be a good approximation.

### B. Example: directional Kalman filter

The tools presented thus far allow us to build a model akin to the Kalman filter [15] over the unit sphere. Previous algorithms for Riemannian Kalman filters predominantly rely on unscented transforms to adapt to the nonlinear state space [13, 17]. Let $\mathbf{y}_t$ be the unobserved variable at time $t$ with initial density $\mathbf{y}_0 \sim \mathsf{SN}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ and assume the predictive distribution

$$p(\mathbf{y}_{t+1} \mid \mathbf{y}_t) = \mathsf{SN}\left(\mathbf{y}_{t+1} \;\middle|\; \mathbf{R}\boldsymbol{\mu},\, \mathbf{R}\boldsymbol{\Lambda}_0\mathbf{R}^\intercal + [\boldsymbol{\Lambda}_{\text{pred}}]_{\boldsymbol{\mu}}\right), \tag{26}$$

where $\mathbf{R}$ is a rotation matrix encoding the expected temporal development and $[\boldsymbol{\Lambda}_{\text{pred}}]_{\boldsymbol{\mu}}$ denotes the concentration of the predictive distribution expressed in the basis of $\mathcal{T}_{\boldsymbol{\mu}}$. Assuming a spherically normal likelihood, the suggested Laplace approximation can be used to estimate the posterior.

We implement this for tracking the direction of the left *femur* (thigh bone) of a person walking [12]. Figure 7B shows the original data (blue), the spherical normal filtered path (orange), and a von Mises-Fisher filtered path (purple). We see that the spherical normal filter smooths the observed data as expected, while the von Mises-Fisher filter appears to have a bias that takes the filtered path outside the data support.

### C. Unknown concentration, known mean

Following standard approaches, we here assume a Wishart prior $p(\boldsymbol{\Lambda} \mid \boldsymbol{\Lambda}_0, m) = \mathcal{W}(\boldsymbol{\Lambda} \mid \boldsymbol{\Lambda}_0, m)$ for the concentration, where we assume the precision is expressed with respect to an orthogonal basis of $\mathcal{T}_{\boldsymbol{\mu}}$. As before, we note that this prior is not conjugate to the spherical normal, and again employ a Laplace approximation of the posterior. The sample space of the posterior is the cone of positive definite matrices, so the Laplace approximation need to be adapted to this constraint.

The log-posterior in this setting is

$$\begin{aligned} h(\boldsymbol{\Lambda}) = &\tfrac{m-(D-1)-1}{2} \log\det(\boldsymbol{\Lambda}) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\Lambda}\right) \\ &- \frac{1}{2}\sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^\intercal \boldsymbol{\Lambda} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n) - N\log(\mathcal{Z}_2(\boldsymbol{\Lambda})). \end{aligned} \tag{27}$$

As in the maximum likelihood setting, we apply the approximate normalization constant (15). The derivative then become

$$\begin{aligned} \frac{\partial h}{\partial \boldsymbol{\Lambda}} = -\frac{1}{2}\Bigg\{ &(D-m)\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Lambda}_0^{-1} \\ &+ \sum_{n=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)^\intercal - aN\,(a\,\boldsymbol{\Lambda} + b\,\mathbf{I})^{-1} \Bigg\}. \end{aligned} \tag{28}$$

This can then be optimized using standard Riemannian optimization [2] to find a maximum a posteriori estimate of $\boldsymbol{\Lambda}$. As for the mean, we can build a Laplace approximation

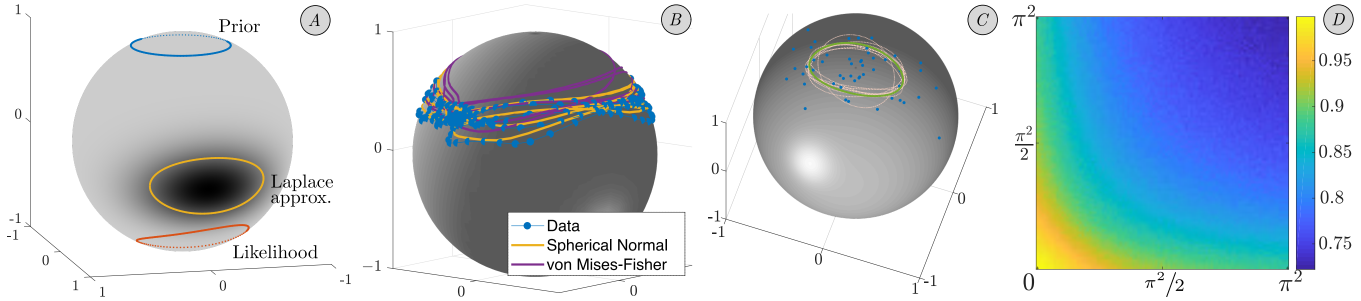

Fig. 7. *A:* Laplace approximation of the mean posterior. *B:* data (blue) and the filtered path (orange). *C:* samples from a Laplace approximation to the concentration posterior. *D:* acceptance rates for the sampling algorithm.

**Algorithm 3** Sampling $\mathbf{x} \sim \mathsf{SN}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$.

```
repeat
    $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{\Lambda} + \frac{D-2}{\pi}\mathbf{I}\right)^{-1}\right)$.
    $r \leftarrow \frac{\exp\left(-\frac{1}{2}\mathbf{v}^\intercal\boldsymbol{\Lambda}\mathbf{v}\right)\left(\frac{\sin(\|\mathbf{v}\|)}{\|\mathbf{v}\|}\right)^{D-2}}{\exp\left(-\frac{1}{2}\mathbf{v}^\intercal\left(\boldsymbol{\Lambda}+\frac{D-2}{\pi}\mathbf{I}\right)\mathbf{v}\right)}$.
    $u \sim \mathsf{Uniform}(0, 1)$.
until $\|\mathbf{v}\| \leq \pi$ and $u \leq r$.
$\mathbf{x} \leftarrow \mathrm{Exp}_{\boldsymbol{\mu}}(\mathbf{v})$.
```

of the posterior of $\boldsymbol{\Lambda}$. Extending the standard derivation of the Laplace approximation [18] shows that the approximate posterior of $\boldsymbol{\Lambda}$ follow a matrix log-normal distribution over the positive definite cone [24], where the concentration is given by the Riemannian Hessian of the log-posterior at $\boldsymbol{\Lambda}_{\text{MAP}}$. As an example, Fig. 7C shows samples from the Laplace posterior estimated from 50 observations and a Wishart prior. The choice of a Wishart prior is not essential as there are no conjugate properties to exploit.

### D. *Unknown mean and concentration*

The concentration matrix $\boldsymbol{\Lambda}$ is fundamentally tied to the mean $\boldsymbol{\mu}$ since it is expressed with respect to a basis of $\mathcal{T}_{\boldsymbol{\mu}}$. Changing $\boldsymbol{\mu}$, thus, renders $\boldsymbol{\Lambda}$ meaningless. This complicate a joint model of both $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, and we do not investigate the issue further in this manuscript.

## VI. Sampling

In many computational pipelines it is essential to be able to draw samples from the applied distributions. We therefore provide a simple, yet efficient, algorithm for simulating the spherical normal. Expressed in the tangent space of $\boldsymbol{\mu}$ the spherical normal has distribution

$$\mathsf{SN}_{\mathcal{T}_{\boldsymbol{\mu}}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\mathbf{v}^\intercal\boldsymbol{\Lambda}\mathbf{v}\right)\left(\frac{\sin(\|\mathbf{v}\|)}{\|\mathbf{v}\|}\right)^{D-2}. \tag{29}$$

We note that $\exp(-{(D-2)\|\mathbf{v}\|^2}/{2\pi}) \geq \left({\sin(\|\mathbf{v}\|)}/{\|\mathbf{v}\|}\right)^{D-2}$ for $\|\mathbf{v}\| \leq \pi$. Since this envelope is reasonably tight, we propose a basic rejection sampling [7] scheme where tangent vectors are proposed from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda} + {(D-2)}/{\pi}\mathbf{I})$. This is summarized in Algorithm 3 and Fig. 7D shows the acceptance rate for ${1}/{\lambda_1}$ and ${1}/{\lambda_2}$ in the range $(0, \pi^2]$. On average the acceptance rate over this domain is 82%; for large concentrations (the common scenario) the acceptance rate is higher. This is quite efficient.

## VII. Summary and Outlook

This paper contributes the first practical tools for efficient curvature-aware inference over the unit sphere. We provide a closed-form expression for the normalization constant of isotropic spherical normal distributions in any dimension. From this we build good approximations for the anisotropic case in $\mathcal{S}^2$. We provide efficient algorithms for maximum likelihood estimation of the mean and concentration parameters of the distribution, and exemplify these with an EM algorithm for mixtures of spherical normals. We further provide tools for approximate Bayesian inference in the form of Laplace approximations for the posteriors of both the mean and the concentration parameters. We exemplify these approximations with a spherical Kalman filter. Finally, we provide a simple yet efficient sampling algorithm for simulating the spherical normal.

Several questions, however, remain open. Most importantly, our approximation to the anisotropic normalization constant does not extend beyond $\mathcal{S}^2$, which is a strong limitation. This is similar to established models in directional statistics, where the anisotropic Fisher-Bingham distribution only has a known normalization constant for $\mathcal{S}^2$.

The proposed Laplace approximations carry over to other Riemannian manifolds, where Bayesian inference is rarely applied. Perhaps more importantly, the Laplace approximation of the posterior concentration matrix also carry over to the Euclidean domain, where it can be of value when approximating posteriors over precision matrices. The resulting matrix log-normal distribution over the positive definite cone is significantly more informative than the commonly applied Wishart distribution, as it can capture anisotropic distributions over the positive definite cone.

It may be beneficial to consider variational approximations to the posteriors rather than the proposed Laplace approximations. This should be feasible when using the proposed approximate normalization constant.

# REFERENCES

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723, 1974.

[4] S. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Math. Soc., 2007.

[5] G. Arvanitidis, L. K. Hansen, and S. Hauberg. A locally adaptive normal distribution. In *Advances in Neural Information Processing Systems*, 2016.

[6] 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*, 6: 1345–1382, 2005.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[8] M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2011.

[9] M. do Carmo. *Riemannian Geometry*. Mathematics (Boston, Mass.). Birkhäuser, 1992.

[10] T. Hamelryck, J. T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2(9):e131, 2006.

[11] M. A. Hasnat. *Unsupervised 3D image clustering and extension to joint color and depth segmentation*. PhD thesis, Université Jean Monnet - Saint-Etienne, Oct. 2014.

[12] S. Hauberg. Principal Curves on Riemannian Manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[13] S. Hauberg, F. Lauze, and K. S. Pedersen. Unscented kalman filtering on riemannian manifolds. *Journal of Mathematical Imaging and Vision (JMIV)*, 46(1):103–120, 2013.

[14] S. Hauberg, A. Feragen, R. Enficiaud, and M. J. Black. Scalable robust principal component analysis using grassmann averages. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[15] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[16] J. T. Kent. The Fisher-Bingham Distribution on the Sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):71–80, 1982.

[17] G. Kurz, I. Gilitschenski, and U. D. Hanebeck. Recursive nonlinear filtering for angular data based on circular distributions. In *American Control Conference (ACC), 2013*, pages 5439–5445. IEEE, 2013.

[18] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[19] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 1999.

[20] X. Pennec. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision (JMIV)*, 25(1):127–154, 2006.

[21] S. Purkayastha. A rotationally symmetric directional distribution: obtained through maximum likelihood characterization. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 70–83, 1991.

[22] R. E. Røge, K. H. Madsen, M. N. Schmidt, and M. Mørup. Infinite von mises–fisher mixture modeling of whole brain fmri data. *Neural Computation*, 29(10):2712–2741, 2017.

[23] H. Salehian, R. Chakraborty, E. Ofori, D. Vaillancourt, and B. C. Vemuri. An efficient recursive estimator of the Fréchet mean on hypersphere with applications to Medical Image Analysis. In *Math. Foundations of Computational Anatomy*, 2015.

[24] A. Schwartzman. Lognormal distributions and geometric averages of symmetric positive definite matrices. *Int. Stat. Review*, 84(3):456–486, 2016.

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012.

[26] V. Smidl and A. Quinn. On Bayesian principal component analysis. *Computational statistics & data analysis*, 51(9): 4101–4123, 2007.

[27] J. Straub, T. Campbell, J. P. How, and J. W. Fisher III. Small-variance nonparametric clustering on the hypersphere. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] J. Straub, J. Chang, O. Freifeld, and J. W. Fisher III. A Dirichlet Process Mixture Model for Spherical Data. In *International Conference on Artificial Intelligence and Statistics*, 2015.

[29] P. Thomas Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision (IJCV)*, 105(2):171–185, 11 2013.

[30] 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)*, 2014.

[31] M. Zhang and P. Fletcher. Probabilistic Principal Geodesic Analysis. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 1178–1186, 2013.