Abstract—In directional statistics, the von Mises-Fisher (vMF) distribution has long been a mainstay for inference with data on the unit hypersphere. The performance of statistical inference based on the vMF distribution, however, may suffer when there are significant outliers and noise in the data. Based on an analogy of the median as a robust measure of central tendency and its relationship to the Laplace distribution, we propose the spherical Laplace (SL) distribution, a novel probability measure for modelling directional data. In this paper, we study foundational properties of the distribution such as theoretical results on maximum likelihood estimation and a sampling scheme for probabilistic inference. We derive efficient numerical routines for parameter estimation in the absence of closed-form formula. An application of model-based clustering is considered under the finite mixture model framework. Our numerical methods for parameter estimation and clustering are validated using simulated and real data experiments.

Index Terms—directional statistics, spherical Laplace distribution, model-based clustering, robust inference

I. INTRODUCTION

A prominent discipline in modern data science is to learn from data equipped with certain geometric constraints. The program often assumes that data reside on Riemannian manifolds [1] and has attracted much attention over the last decades in both theoretical foundation [2]–[4] and practical applications across many fields including computer vision [5], [6] and medical image analysis [7]–[9]. In statistics, the relevant research area has long been known as directional statistics [10], [11] whose objects of interests include directions, axes, and rotations. Especially, directions seem to take the largest portion in the field for its popularity as a number of applications involve observations on the unit hypersphere, which is a model manifold of constant positive curvature.

An off-the-shelf choice of probability measure on the unit hypersphere is the von Mises-Fisher (vMF) distribution [12], which is a location-scale family distribution that has higher concentration of mass near a mean direction. Recently, the spherical normal (SN) distribution [13] was proposed as an alternative by adopting the geodesic distance on the unit hypersphere. Both distributions are Gaussian-like laws since they are defined using negated squared distances to their location parameters within an exponential function. For both distributions, the maximum likelihood estimate of a location parameter corresponds to the quantity called the Fréchet mean [14], which is a minimizer of sum of squared distances.

While these distributions have played an important role in statistical inference, their performance may deteriorate when there exist outliers or noise of large magnitude. The literature of robust statistics has well acknowledged that the median is a robust measure of central tendency [15]. This has been translated on manifolds to replace the Fréchet mean with the Fréchet median, a minimizer for sum of distances [8], [16]. Hence, it is natural to ask what distribution is related to the concept of Fréchet median on the unit hypersphere as the vMF and SN distributions to the Fréchet mean.

This motivates our paper to propose the spherical Laplace (SL) distribution, which generalizes the Laplace distribution on the real line onto the unit hypersphere by defining its density using a scaled distance to a mean location rather than squared distance. A leading contribution of this paper is to establish the SL distribution and derive an explicit, simplified form of a normalizing constant amicable for numerical approximation. We also consider a rejection-based sampling scheme to draw random samples from the SL distribution.

Another class of our contribution is on parameter estimation in the sense of maximum likelihood, which is one of the fundamental tasks for statistical inference. We show that under mild support conditions the SL distribution admits unique parameter estimates. Since the maximum likelihood estimates are solutions of some nonlinear equations, we present algorithms to estimate the parameters. This is applied to model-based clustering [17] by employing the finite mixture model framework [18] where each component is the SL distribution. Computation for the SL mixture is heavily dependent on a generalized version of the parameter estimation algorithms where observations are given fixed weights. We present computational pipeline in detail for the SL mixture in a similar fashion to its predecessors [19], [20]. Computational apparatus for the SL distribution is available in an R statistical package Riemann (version 0.1.4) [21] available at https://CRAN.R-project.org/package=Riemann. The accompanying Supplementary Material is available at https://github.com/kisungyou/papers containing all proofs and others to be mentioned in the paper.

II. BACKGROUND

We first introduce notations throughout the paper. Let

$$\mathbb{S}^p = \{ x \in \mathbb{R}^{p+1} : \|x\| = 1 \}$$

be a p-dimensional unit sphere...
where the exponential map is well-defined in a geodesic ball where the exponential map is well-defined. In general, an exponential map is only locally defined. The logarithmic map $\log$ is reserved to denote the geodesic distance $d(x, y)$.

Fig. 1: Visual representation of exponential (red) and logarithmic (blue) maps on a Riemannian manifold $M$ for some points $x, y \in M$ and a tangent vector $u \in T_x M$.

### A. Geometry of hypersphere

The unit hypersphere is a model space for a simply connected manifold of constant positive curvature [1], [22]. We review basic properties of the unit sphere as a Riemannian manifold and explicit forms of computational operations [23].

For some $x \in S^p$, the tangent space is characterized as $T_x S^p = \{ z \in \mathbb{R}^{p+1} | \langle x, z \rangle = 0 \}$ with a standard inner product $\langle \cdot, \cdot \rangle$. The Riemannian metric for $S^p$ is a canonical metric, i.e., $g_x(u, v) = \langle u, v \rangle$ for all $u, v \in T_x S^p$. The shortest path connecting any two points $x, y \in S^p$ is along the great circle in that the geodesic distance is given by $d(x, y) = \arccos(\langle x, y \rangle)$. Given a pair $(x, u) \in M \times T_x M$ for some Riemannian manifold $M$, there exists a unique geodesic $\gamma : [0, 1] \to M$ such that $\gamma(0) = x$ and $\gamma'(0) = u$. An exponential map is defined by $\exp_x(u) = \gamma(1)$, which maps a vector in some tangent space to the manifold itself whose inverse is called a logarithmic map, as shown in Fig. 1. When $M = S^p$, two maps are endowed with explicit formula. Define an operator $\text{Proj}_x(z) = z - \langle z, x \rangle x$ that projects some vector $z \in \mathbb{R}^{p+1}$ onto the tangent space $T_x S^p$. For $x, y \in S^p$ and $u \in T_x S^p$, the exponential map $\exp_x(\cdot) : T_x S^p \to S^p$ and the logarithmic map $\log_x(\cdot) : S^p \to T_x S^p$ are expressed as

$$\exp_x(u) = \cos(||u||)x + \frac{\sin(||u||)}{||u||}u, \quad (1a)$$

$$\log_x(y) = \frac{d(x, y)}{||\text{Proj}_x(y - x)||} \text{Proj}_x(y - x). \quad (1b)$$

In general, an exponential map is only locally defined. The injectivity radius $\text{inj}(M)$ of a manifold $M$ is a maximal radius of a geodesic ball where the exponential map is well-defined and a diffeomorphism, which is $\pi$ for the unit hypersphere.

### B. The von Mises-Fisher and spherical normal distributions

On $S^p$, the vMF distribution is one of the most popular probability laws that has seen many applications [19], [24], [25]. Parametrized by location $\mu \in S^p$ and scale $\kappa \in \mathbb{R}^+$ parameters, the vMF distribution has the following density function of the form $f_{\text{vMF}}(x | \mu, \kappa) \propto \exp(\kappa x^T \mu)$. It was addressed in [26] that $\exp(\kappa x^T \mu) \propto \exp(-\kappa ||x - \mu||^2/2)$, indicating that the vMF distribution can be roughly considered as a direct adaptation of the isotropic Gaussian distribution onto the unit hypersphere.

It is natural to question whether the use of standard Euclidean distance in defining a probability measure on the constrained domain of hypersphere is appropriate. This led to a recent proposal of the SN distribution to replace the Euclidean norm with the geodesic distance. For an isotropic version of the SN distribution, the density function, parametrized by location $\mu \in S^p$ and concentration $\lambda \in \mathbb{R}^+$ parameters, is written as $f_{\text{SN}}(x | \mu, \lambda) \propto \exp(-\lambda \cdot d^2(x, \mu)/2)$.

It is worth to mention the connection between the vMF and SN distributions where both laws are parametrized by scaled square of distances from a location parameter. In the literature of statistics on manifolds, these two formulations correspond to the dichotomy of intrinsic and extrinsic frameworks [3]. The intrinsic approach measures dissimilarity of two points on a manifold by the shortest-path geodesic joining the two, while the extrinsic framework uses a standard norm after an equivariant embedding of the data on manifold onto the Euclidean space [27]. Therefore, the vMF and SN distributions correspond to extrinsic and intrinsic Gaussian-like probability laws on the sphere, respectively.

### III. The spherical Laplace distribution

#### A. Definition

We propose the spherical Laplace (SL) distribution, which is an isotropic location-scale family distribution on the unit hypersphere $S^p$ of $p \geq 1$.

**Definition 1.** Let $\mu \in S^p$ and $\sigma \in \mathbb{R}^+$ be location and scale parameters, respectively. Density function of the SL distribution is defined by

$$f_{\text{SL}}(x | \mu, \sigma) = \frac{1}{C_p(\mu, \sigma)} \exp\left(-\frac{d(x, \mu)}{\sigma}\right),$$

for a normalizing constant $C_p(\mu, \sigma)$,

$$C_p(\mu, \sigma) = \int_{S^p} \exp\left(-\frac{d(x, \mu)}{\sigma}\right) dx. \quad (2)$$

The scale parameter quantifies the degree of dispersion as shown in Fig. 2, where a small $\sigma$ leads to concentrated mass around $\mu$ and a large $\sigma$ leads to largely dispersed distribution of the mass.

It is trivial to observe that (2) is well defined. The integrand is a smooth, bounded function due to the fact that $\sigma$ is strictly positive and $d(x, \mu)$ is bounded in $(0, \pi)$. Since the domain is compact, the integral is finite. However, evaluation of (2) involves multidimensional integration over the constrained
domain, which makes practical usage of the distribution prohibitive. Although a closed-form formula is not available, we show in the following proposition that the expression for normalizing constant can be transformed into a single-variable integration over the bounded domain.

**Proposition 1.** The normalizing constant (2) can be written as a univariate integral as

$$C_p(\mu, \sigma) = A_{p-1} \int_{r=0}^{\pi} \exp\left(-\frac{r}{\sigma}\right) \sin^{p-1}(r) dr,$$

where $A_{p-1} = 2\pi^{p/2}/\Gamma(p/2)$ is the hypervolume or surface area of $S^{p-1}$ and $\Gamma(\cdot)$ is the standard Gamma function.

Proposition 1 does not only lower computational complexity to evaluate the normalizing constant, but also implies that the normalizing constant is not contingent on the choice of $\mu$. In what follows, we will denote the normalizing constant of the SL distribution as $C_p(\sigma)$ to emphasize its independence on a choice of location $\mu$.

**B. Sampling**

It is often of fundamental interest to draw random samples for a given probability distribution. We start by employing a standard rejection sampling technique for its ease of use and efficiency [28]. We take the SN distribution as a proposal density for the rejection sampler whose sampling strategy was well studied in [13]. For a SL distribution with parameters $(\mu, \sigma) \in S^p \times \mathbb{R}^+$, setting a SN concentration parameter as $\lambda = 1/\sigma$ with the same location $\mu$ reduces to acceptance-rejection, which is summarized in Algorithm 1. The sampling scheme is described in detail in the Supplementary Material.

While the rejection sampler is a convenient choice to sample from the SL distribution, one drawback is that the efficiency deteriorates when $\sigma \approx 0$. Recall that the distance is upper bounded by $\pi$ for any two points on the unit hypersphere. When a draw $y$ from a corresponding SN distribution is given, the acceptance threshold $\tau$ is defined as

$$\tau = \exp\left(\frac{(d(\mu, y) - 1)^2 - (\pi - 1)^2}{2\sigma}\right),$$

where the numerator within an exponential function is always smaller than or equal to 0. It is trivial to observe that $\tau \to 1$ as $\sigma \to \infty$, implying that the rejection sampler almost always accepts a new sample for larger values of a scale parameter $\sigma$. On the other hand, when $\sigma \to 0$, the threshold converges to 0 in that the sampler requires a tremendously larger number of repeated sampling from its proposal distribution. Hence, it is suggested to use other Markov chain Monte Carlo methods such as the Metropolis-Hastings (MH) algorithm [29], [30] in the regime of small $\sigma$ values. In our experiments, we used the MH sampler where a sequence of samples is generated from a Gaussian distribution on the tangent space at an iterate.

**C. Maximum likelihood estimation**

We now consider the task of maximum likelihood estimation for parameters of the SL distribution. Let $x_1, \ldots, x_N \in S^p$ be an i.i.d sample from the SL distribution with two parameters, location $\mu$ and scale $\sigma$. The maximum likelihood estimates $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}) \in S^p \times \mathbb{R}^+$ are maximizers of the following log-likelihood function

$$L(\mu, \sigma) = -\frac{1}{\sigma} \sum_{n=1}^{N} d(x_n, \mu) - N \log C_p(\sigma).$$

One can observe that (4) does not admit closed-form formulae of its maximizers because of the nonlinear terms, engendering the need to employ computational approach for parameter estimation, which will be discussed in the next section.

A closer look at (4) allows us to easily recognize that the constrained maximization with respect to $\mu$ is independent of $\sigma$ since $\mu_{MLE} = \arg\max_{\mu} -\sum_{n=1}^{N} d(x_n, \mu)/\sigma$ for $\mu \in S^p$, which is equivalent to $\arg\min_{\mu} \sum_{n=1}^{N} d(x_n, \mu)$. The literature of manifold-valued statistics has recognized this object as the Fréchet or geometric median problem [8], [16]. Once $\mu_{MLE}$ is obtained, solving for $\sigma$ is merely a univariate minimization problem, i.e., $\sigma_{MLE} = \arg\min_{\sigma} S/\sigma + \log C_p(\sigma)$ for some $\sigma \in \mathbb{R}^+$ and a fixed constant $S = \sum_{n=1}^{N} d(x_n, \mu_{MLE})/N$. We show existence and uniqueness of maximum likelihood estimates in the following theorem under some conditions.

**Theorem 2.** Let $x_1, \ldots, x_N$ be an i.i.d sample on a $p$-dimensional unit hypersphere $S^p$. If the sample is contained in an open geodesic ball $B(x, \pi/4)$ for some $x \in S^p$ and not totally contained in any geodesic, maximum likelihood estimates $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE})$ uniquely exist.
IV. PARAMETER ESTIMATION

It was observed that the maximum likelihood estimation can be decomposed into two components in a sequential manner to solve the standard Fréchet median problem for the location parameter $\mu \in \mathbb{S}^p$ and a univariate optimization problem for the scale parameter $\sigma \in \mathbb{R}^+$. Throughout this section, we consider a more general scenario where each $x_n$ is weighted by a non-negative constant $w_n > 0$ that sums to 1, i.e., $\sum_{n=1}^N w_n = 1$, and present adapted algorithms to the context for both maximum likelihood estimation and numerical apparatus for model-based clustering in the following section.

A. Estimation of the location

Given a random sample $x_1, \ldots, x_N \in \mathbb{S}^p$ and weights $w_1, \ldots, w_N$, the weighted Fréchet median problem is to solve

$$\min_{\mu \in \mathbb{S}^p} F(\mu) = \min_{\mu \in \mathbb{S}^p} \sum_{n=1}^N w_n d(x_n, \mu), \quad (5)$$

where $\hat{\mu}_{\text{MLE}}$ is a solution to the special case of (5) if all weights $w_n$'s are equal such as $w_1 = \cdots = w_N = 1/N$.

We employ a geometric variant of the Weiszfeld algorithm [8], whose standard version has long been known in the Euclidean regime [31]. In general, the Weiszfeld algorithm follows a descent path, making it to belong to a class of gradient descent methods. At each iteration, an iterate is updated by varying amount of step size determined by a fixed rule. This removes potentially large computational burden to tune step sizes by line search at each iteration.

Now a geometric Weiszfeld algorithm is described for the weighted Fréchet median problem, which is summarized in Algorithm 2. Roughly speaking, the gradient $\nabla F|_x$ of a scalar-valued function $F$ on a manifold $M$ is defined as a vector field, meaning that it takes a value as a tangent vector in $T_xM$. Once an update is done in the gradient direction, an additional step is necessary to push an iterate onto the manifold itself [23]. Using the variation of energy, it can be readily shown that $\nabla d(\mu, x) = -\log \mu(x)/d(\mu, x)$, leading to define a gradient of (5) at iteration $t$ as

$$\nabla F(\mu(t)) = -\sum_{n=1}^N \frac{w_n}{d(x_n, \mu(t))} \log \mu(t)(x_n).$$

The geometric Weiszfeld algorithm updates an iterate by $\mu^{(t+1)} = \text{Exp}_{\mu(t)}(-\alpha(t)\nabla F(\mu(t)))$ for $\alpha(t) = 1/(\sum_{n=1}^N w_n)^2$ that is automatically determined at each iteration with a scaled weight $w_n(t) = w_n/d(x_n, \mu(t))$. This leads to the update rule

$$\mu^{(t+1)} = \text{Exp}_{\mu(t)}\left(\frac{\sum_{n=1}^N w_n(t) \log \mu(t)(x_n)}{\sum_{n=1}^N w_n(t)}\right), \quad (6)$$

where the exponential and logarithmic maps are computed using (1a) and (1b).

In the initial stage of Algorithm 2, a starting point is taken as $\mu^{(0)} = \sum_{n=1}^N w_n x_n/\sum_{n=1}^N w_n x_n$. This quantity is equivalent to an extrinsic mean of a random sample [3], which is also a maximum likelihood estimate of a location parameter from the vMF distribution. When a random sample is adequately concentrated as stated in Theorem 2, an extrinsic mean is a reasonable choice as it is contained in a convex geodesic ball while it does not require expensive computational routines. Next, it is standard to halt iterations when an iterate belongs to a set of given observations for the Weiszfeld algorithm since a scaled weight cannot be evaluated if $d(\mu(t+1), x_n) = 0$ for any $n$ at the next iteration. Ad hoc remedies include to re-initialize an algorithm, remove a matching point from computation in the specific step, or pad 0 with a very small positive number, all of which we do not consider in this paper.

B. Estimation of the scale

Maximum likelihood estimation for the scale parameter $\sigma$ of the SL distribution reduces to a univariate minimization of the objective function $g(\sigma) = S/\sigma + \log C_p(\sigma)$ given a fixed constant $S = \sum_{n=1}^N d(x_n, \mu_{\text{MLE}})/N$. Theorem 2 states that the cost function admits a unique critical point. Hence, it can be cast as a root-finding problem of $g'(\sigma) = 0$.

For scale parameter estimation, we use the Newton-Raphson method [32], [33]. When a univariate function of interest, say $f$, has continuous derivatives up to order 0, the Newton-Raphson algorithm solves $f(x) = 0$ by updating an iterate as $x^{(t+1)} = x^{(t)} - f(x^{(t)})/f'(x^{(t)})$ with an initial point $x^{(0)}$. By setting $f = g'$, we can derive the updating rule to solve $g'(\sigma) = 0$ by $\sigma^{(t+1)} = \sigma^{(t)} - g''(\sigma^{(t)})/g'(\sigma^{(t)})$. Define

$$I_0(\sigma) = \int_{r=0}^\infty \exp \left(-\frac{r}{\sigma}\right) \sin^{p-1}(r) dr,$$

$$I_1(\sigma) = \int_{r=0}^\infty \frac{r}{\sigma^2} \exp \left(-\frac{r}{\sigma}\right) \sin^{p-1}(r) dr,$$

$$I_2(\sigma) = \int_{r=0}^\infty \frac{r^2}{\sigma^4} \exp \left(-\frac{r}{\sigma}\right) \sin^{p-1}(r) dr,$$

all of which are integrals related to the normalizing constant and its derivatives. By re-writing derivatives of $g$ with respect to the above integrals, we can simplify the Newton-Raphson update at iteration $t$

$$\sigma^{(t+1)} = \sigma^{(t)} - A^{(t)}/B^{(t)}, \quad (7)$$

Algorithm 2 Weighted Fréchet median computation

Input: a random sample $x_1, \ldots, x_N \in \mathbb{S}^p$, weights $w_1, \ldots, w_N$, stopping criterion $\epsilon$.

Output: $\hat{\mu} = \text{argmin} \sum_{n=1}^N w_n d(x_n, \mu)$ for $\mu \in \mathbb{S}^p$.

Initialize $\mu^{(0)} = \sum_{n=1}^N w_n x_n/\sum_{n=1}^N w_n x_n$.

repeat

Compute $w_n(t) = w_n/d(\mu(t), x_n)$ for $n \in \{1, \ldots, N\}$. Update an iterate $\mu^{(t+1)}$ by (6).

if $\mu^{(t+1)}$ equals to one of $x_1, \ldots, x_N$ then stop the algorithm.

end if

until $d(\mu(t), \mu^{(t+1)}) < \epsilon$.

Take $\hat{\mu} = \mu^{(t+1)}$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Algorithm 3 Maximum likelihood estimation of the scale parameter

**Input:** a random sample \(x_1, \ldots, x_N \in \mathbb{R}^p\),
a constant \(S\), stopping criterion \(\epsilon\).

**Output:** \(\hat{\sigma}_{\text{MLE}} = \arg\min \frac{S}{\sigma} + \log C_p(\sigma) \) for \(\sigma \in \mathbb{R}^+\).

**Initialize** \(\sigma^{(0)}\).

**repeat**

Update an iterate \(\sigma^{(t+1)}\) by either (7) or (8).

**until** \(|\sigma^{(t)} - \sigma^{(t+1)}| < \epsilon\).

Take \(\hat{\sigma}_{\text{MLE}} = \sigma^{(t+1)}\).

for the two terms \(A^{(t)}_\sigma\) and \(B^{(t)}_\sigma\) are defined as

\[
A^{(t)}_\sigma = \left(-\frac{S}{(\sigma^{(t)})^2} + \frac{I_1(\sigma^{(t)})}{I_0(\sigma^{(t)})}\right),
\]

\[
B^{(t)}_\sigma = \left(\frac{2S}{(\sigma^{(t)})^2} + \frac{I_0(\sigma^{(t)})I_2(\sigma^{(t)}) - (I_1(\sigma^{(t)}))^2}{(I_0(\sigma^{(t)})^2}\right).
\]

When \(\sigma\) is close to zero, evaluation of \(I_1(\sigma)\) and \(I_2(\sigma)\) may suffer from numerical instability due to presence of \(\sigma\) in rational terms. As a result, an alternative to (7) is considered by approximating derivatives with the centered finite difference schemes so that we have

\[
\sigma^{(t+1)} = \sigma^{(t)} - \frac{h(g(\sigma^{(t)} + h) - g(\sigma^{(t)} - h))}{2(g(\sigma^{(t)} + h) - 2g(\sigma^{(t)}) + g(\sigma^{(t)} - h))},
\]

for a sufficiently small step-size \(h > 0\). We note that an approximate updating rule (8) has equivalent computational complexity to that of (7) as both integrate one-dimensional smooth functions three times per iteration.

V. APPLICATION TO CLUSTERING

One of the direct applications for parameter estimation of the SL distribution is probabilistic clustering of sphere-valued data using finite mixture models [18]. The density of a finite mixture model of \(K\) SL components is

\[
h(x|\Theta) = \sum_{k=1}^{K} \pi_k f_{\text{SL}}(x|\mu_k, \sigma_k),
\]

for component weights \(\pi_k, k \in \{1, \ldots, K\}\) such that \(\sum_{k=1}^{K} \pi_k = 1\) and parameters \(\Theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}\). We follow a standard approach to maximize log-likelihood given a random sample \(X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^p\) by introducing latent variables for class membership and applying Expectation-Maximization (EM) algorithm [34]. We refer interested readers to [28] for thorough description of the technique. To briefly introduce, denote a binary matrix of latent class memberships as \(Z \in \{0, 1\}^{N \times K}\) in that every row of \(Z\) contains only one non-zero element and the joint distribution of \(X\) and \(Z\) as

\[
P(Z, X | \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k f_{\text{SL}}(x_n | \mu_k, \sigma_k))^{z_{nk}}.
\]

The EM algorithm alternates E- and M-steps. At iteration \(t\), the E-step evaluates posterior distribution of the latent variable \(P(Z | X, \Theta^{(t)})\) to compute the complete-data log-likelihood \(Q(\Theta | \Theta^{(t)}) = E_{Z|X,\Theta^{(t)}}[\log P(X, Z | \Theta)]\) and the M-step updates all parameters \(\Theta^{(t+1)}\) by maximizing \(Q(\Theta | \Theta^{(t)})\).

We now elaborate on explicit expressions for updating parameters in a finite mixture of SL distributions under the standard EM framework at iteration \(t\). The E-step starts by evaluating posterior of the latent variables, which first reduces to evaluating \(E[z_{nk}]\) by

\[
\gamma_{nk} := E[z_{nk}] = \frac{\pi_k f_{\text{SL}}(x_n | \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\text{SL}}(x_n | \mu_j, \sigma_j)}.
\]

for \((n, k) \in \{1, \ldots, N\} \times \{1, \ldots, K\}\). Let \(\Gamma := \gamma_{nk} \in [0, 1]^{N \times K}\) denote a matrix of posterior for the latent variables known as soft clustering or membership matrix. An \((n, k)\)-th entry of \(\Gamma\) encodes information on how likely an observation \(x_n\) belongs to the \(k\)-th cluster. Given the evaluations, the complete-data log-likelihood can be written as

\[
Q(\Theta | \Theta^{(t)}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\{\log \pi_k + \log f_{\text{SL}}(x_n | \mu_k, \sigma_k)\right\},
\]

which is an objective function for maximization in the M-step. For each component \(k \in \{1, \ldots, K\}\), the weights are updated by \(\pi_k^{(t+1)} = \sum_{n=1}^{N} \gamma_{nk}/N\). The new location parameters are solutions of the following weighted Fréchet median problems,

\[
\mu_k^{(t+1)} = \arg\min_{\mu \in \mathbb{R}^p} \sum_{n=1}^{N} \gamma_{nk} \cdot d(x_n, \mu),
\]

which can be solved using Algorithm 2 with \(w_i = \gamma_{ik}, i \in \{1, \ldots, N\}\). Lastly, individual scale parameters are solutions of the following problems,

\[
\sigma_k^{(t+1)} = \arg\min_{\sigma \in \mathbb{R}^+} \left(\frac{\sum_{n=1}^{N} \gamma_{nk} \cdot d(x_n, \mu_k^{(t+1)})}{\sigma \cdot \sum_{n, k} \gamma_{nk}}\right) + \log C_p(\sigma),
\]

which can be solved by Algorithm 3. The EM algorithm for our mixture model is described in the Supplementary Material.

We now turn to discuss some aspects of the model and EM algorithm. First, one way to regularize varying scales across multiple components is to use a common scale parameter rather than component-specific values, i.e., \(\sigma = \sigma_1 = \cdots = \sigma_K\). In this setting, the model is called homogeneous and a common scale parameter is updated by

\[
\sigma^{(t+1)} = \arg\min_{\sigma \in \mathbb{R}^+} \left(\frac{\sum_{n=1}^{N} \gamma_{nk} \cdot d(x_n, \mu_k^{(t+1)})}{\sigma \cdot \sum_{n, k} \gamma_{nk}}\right) + \log C_p(\sigma).
\]

Second, computational cost of the algorithm becomes prohibitive when the sample size \(N\) grows, mainly due to the weighted Fréchet median problem. We present two popular heuristics to reduce the sample size - hard and stochastic assignments - that directly manipulate \(\Gamma\) so that the pertained computation is limited to a smaller subset of the sample. Denote the \(n\)-th row of \(\Gamma\) as \(\Gamma_n\) and sample \((\nu_z, p = \nu_z)\) is a sampling procedure to randomly draw an element from
a vector $v_z$ with probability $p_z$. Two heuristic assignments, hard and stochastic, are defined as

$$\text{hard}(\gamma_{nk}) = 1\{k \text{ is maximal index of } \Gamma_n\}, \quad (12)$$
$$\text{stochastic}(\gamma_{nk}) = 1\{k = \text{sample}(1 : K, p = \Gamma_n)\}, \quad (13)$$

for an indicator function $1\{\cdot\}$. Note that the hard assignment is equivalent to predict discrete-valued label for an observation, while the stochastic assignment performs the same task in a probabilistic manner. Given a large sample, these sparsification strategies help to reduce execution time and space complexity in updating both location and scale parameters. The hard assignment is optimal in the sense that a lower bound of the incomplete-data log-likelihood is maximized [19].

VI. EXPERIMENTS

We come to assess how the proposed algorithms perform for the tasks of parameter estimation and model-based clustering. The first experiment is composed of evaluating parameter estimation algorithms with simulated data. Then, we investigate effectiveness of model-based clustering with the finite mixture of SL distributions using simulated and real data examples.

A. Estimating parameters

We evaluate performance of the proposed algorithms for maximum likelihood estimation of two parameters. We start by presenting the simplest case where two fixed parameters $\mu_0 \in (1, 0, 0, 0, 0, 0)^T \in S^5$ and $\sigma_0 = 0.1$ are used to generate random samples of varying size from 25 to 475. Each setting is run 100 times and empirical distributions of two performance measures, run time and accuracy, are reported. For numerical optimization, the stopping criterion of $\epsilon = 10^{-8}$ was used throughout all computations, which is approximately equal to a square root of the machine epsilon in double precision.

For the location estimation problem, we compared the geometric Weiszfeld method as shown in Algorithm 2 and the standard version of Riemannian gradient descent (RGD) algorithm. Given a random sample $X = \{x_1, \ldots, x_n\} \subset S^5$ and a maximum likelihood estimate $\hat{\mu}_{MLE}$, accuracy is measured by the geodesic distance $d(\hat{\mu}_{MLE}, \mu_0) = \cos^{-1}(\langle \hat{\mu}_{MLE}, \mu_0 \rangle)$. Two performance measures under the setting are shown in Fig. 3. As the sample size increases, better estimates are obtained in both methods at almost parallel levels of accuracy. In terms of computational cost, however, the Weiszfeld algorithm consistently outperforms the RGD and the degree of disparity gets larger along the growing sample size. This is an expected phenomenon in the sense that line search of RGD requires repeated evaluation of the cost functional while the Weiszfeld algorithm requires no such procedure. Therefore, the RGD may converge in fewer iterations than the Weiszfeld algorithm at the increased cost of overall computation.

Similarly, we compared performance of the proposed Newton’s methods against two univariate optimization routines that do not require derivative information, whose performance measures are summarized in Fig. 4. Our first choice is a default optimization routine in R programming language [35] that uses golden-section search with successive parabolic interpolation, which we will denote as Roptim. Another is the differential evolution algorithm for global optimization using a heuristic approach [36] and will be denoted as DE. Exact and approximate updating rules for the Newton’s methods are denoted as NewtonE and NewtonA, respectively. We measured accuracy of the estimate $\hat{\sigma}_{MLE}$ by relative error $|\hat{\sigma}_{MLE} - \sigma_0|/\sigma_0$. The quality of estimation was comparable across all methods, which is not surprising since all algorithms start from an identical location estimate. When it comes to computational efficiency, however, both our proposed algorithms show superior performance to the others.

We performed extensive experiments to compare performance of the proposed algorithms across different settings. From the results reported in the Supplementary Material, we observed similar patterns across heterogeneous experimental conditions regarding the accuracy and elapsed time.

B. Clustering

We consider two examples for model-based clustering of sphere-valued data. In both examples, we compare mixtures of SL distributions with soft (MoSL-soft) and hard (MoSL-hard) assignment with a number of competing algorithms including $k$-means (KMEANS) [37], spherical $k$-means (SKMEANS) [38], mixture of vMF distributions (MOVMF) [19], and mixture of SN distributions (MoSN) [26]. In order to evaluate quality of algorithms, three clustering comparison indices between a true or desired label and an attained label are used, including Jaccard index [39], Rand index [40], and normalized mutual information (NMI) [41]. All three measures are valued in $[0, 1]$, quantifying the degree of similarity.

Fig. 3: Performance measures, (a) accuracy and (b) elapsed time, for the location estimation experiment with 100 repeats.

Fig. 4: Performance measures, (a) accuracy and (b) elapsed time, for the scale estimation experiment with 100 repeats.
between two clusterings in a way that the higher an index value is, the more agreement is present.

First, we use simulated data according to the small-mix example from [19] with minor modification as adopted in [26]. The data generating model in this example is a mixture of two SN distributions on $\mathbb{S}^1 \subset \mathbb{R}^2$. Two equally weighted components are parametrized by $(\mu_1, \lambda_1) = ([-0.251, -0.968], 10)$ and $(\mu_2, \lambda_2) = ([0.399, 0.917], 2)$ with heterogeneous dispersions. In each run, a total of 200 observations are randomly drawn from the model. A random sample with true label and clustering results are shown in Fig. 5, where the true label is perfectly recovered from our proposed mixture of SL distributions while the spherical $k$-means and vMF mixture models misclassified observations near the boundary. We repeated the test 100 times and compared average clustering indices, which are summarized in Table 1 in the Supplementary Material. When the number of clusters is set as $K = 2$, we witnessed an interesting phenomenon that the SL mixture is the best performing method even though the data was generated from the mixture of SN components. One explanation is related to robust estimation of cluster means for the SL distribution so that the perturbed observations at the boundary of two components less affect the overall identification of clusters. When the models are misspecified, the SL mixture returned smaller indices than others. One may consider this as an encouraging pattern since a valid mixture model should discourage misspecification given the unknown reality.

We now turn to the real data example using household data, which is a survey data of household-level expenditures on several commodity groups among 40 individuals (20 males, 20 females) [42]. Following the convention of [20], we extracted expenditures of three categories - food, housing, and service. Our interest is on the proportion of each category so that a profile vector is obtained by $\ell_1$ normalization and projected onto $\mathbb{S}^3$ by square-root transformation, which is a common practice in studying compositional data [43]. Quality of clustering performance was measured by regarding the gender label as a ground-truth cluster label so $K = 2$. Results are summarized in Table 2 in the Supplementary Material. Our proposed SL mixture model showed perfect clustering results when model was correctly specified while others show somewhat strange patterns as shown in Fig. 6. This may come from the fact that these algorithms do not acknowledge much for the constrained geometry of the domain. This pattern is also witnessed in two mixture models of the vMF and SN components that the distinct classes are seemingly contaminated by farthest points. This may be caused by the fact that data from the male group is much more dispersed than that of the female group. When the SL distribution was fitted separately, the females’ expenditure data was distributed with $\hat{\sigma}_{\text{MLE}} = 0.0643$, which is much smaller than that of the male group with $\hat{\sigma}_{\text{MLE}} = 0.1426$, which indicates that the male group shows higher degree of heterogeneity. In reality, it is nearly impenetrable to distinguish presence of outliers from noise of substantial magnitude. Therefore, specifying a mixture model with robust components can help to deal with the scenario that we observed.

VII. CONCLUSION

We proposed the SL distribution, a generalization of the Laplace distribution onto the unit hypersphere. The distribution was shown to be well-defined and provided with a sampling scheme, which is a fundamental tool arising in many computational pipelines. We showed that the maximum likelihood estimates of the governing parameters uniquely exist under mild conditions on the support of an empirical measure. We also proposed numerical optimization routines for maximum likelihood estimation of location and scale parameters since the SL distribution does not admit closed-form formulae for the estimates. The proposed algorithm is validated with extensive experiments using simulated data. An application of our proposal is model-based clustering of spherical data under the finite mixture framework where the components are SL distributions. Our experiments with simulated and real data showed that there is gain from the SL mixture where a large amount of noise is suspected.

We close by discussing several topics of interests for future studies. As noted before, the current sampling scheme built
upon rejection sampler could be inefficient when a scale parameter is set to be very small. We expect that further investigation along this direction makes the proposed distribution a more appealing component not only for statistical inference but also for probabilistic deep learning such as variational autoencoders. Another line of research is reserved for refining clustering procedures. While our current exploration focuses on the finite mixture framework, availability of an effective sampler helps to open up opportunities such as a nonparametric Bayesian density estimation [24] and modelling hierarchical and temporal data on the unit hypersphere [25]. The last direction is anisotropic extension of the SL distribution as done in [13] for the SN distribution, which will be beneficial in flexible modelling of directionally dependent data on the unit hypersphere.

REFERENCES

[1] J. M. Lee, Riemannian Manifolds: An Introduction to Curvature. No. 176 in Graduate Texts in Mathematics, New York: Springer, 1997.
[2] X. Pennec, “Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements,” Journal of Mathematical Imaging and Vision, vol. 25, pp. 127–154, July 2006.
[3] A. Bhattacharya and R. Bhattacharya, Nonparametric Inference on Manifolds: With Applications to Shape Spaces. Cambridge: Cambridge University Press, 2012.
[4] V. Patrangenaru and L. Ellingson, Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis. Boca Raton: CRC Press, Taylor & Francis Group, 2016.
[5] A. Bassacco, A. Chiuso, Y. Ma, and S. Sotto, “Recognition of human gaits,” in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 2, (Kauai, HI, USA), pp. II–52–II–57, IEEE Comput. Soc, 2001.
[6] G. Aggarwal, A. Chowdhury, and R. Chellappa, “A system identification approach for video-based face recognition,” in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., (Cambridge, UK), pp. 175–178 Vol.4, IEEE, 2004.
[7] X. Pennec, S. Sommer, and T. Fletcher, Riemannian Geometry in Medical Image Analysis. San Diego: Academic Press, 2020.
[8] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi, “The geometric median on Riemannian manifolds with application to robust atlas estimation,” NeuroImage, vol. 45, pp. S145–S152, Mar. 2009.
[9] K. You and H.-J. Park, “Re-visited Riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity,” NeuroImage, vol. 225, p. 117464, Jan. 2021.
[10] K. V. Mardia and P. E. Jupp, Directional Statistics. Wiley Series in Probability and Statistics, Chichester ; New York: J. Wiley, 2000.
[11] C. Ley and T. Verdebout, Modern Directional Statistics. Chapman and Hall/CRC, first ed., Aug. 2017.
[12] R. Fisher, “Dispersion on a Sphere,” Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 217, pp. 295–305, May 1953.
[13] S. Hauberg, “Directional Statistics with the Spherical Normal Distribution,” in Proceedings of the 21st International Conference on Information Fusion (FUSION), (Cambridge, United Kingdom), pp. 704–711, IEEE, July 2018.
[14] W. S. Kendall, “Probability, Convexity, and Harmonic Maps with Small Image I: Uniqueness and Fine Existence,” Proceedings of the London Mathematical Society, s3-61, pp. 371–406, Sept. 1990.
[15] P. J. Huber, Robust Statistics. Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1981.
[16] M. Arnaudon, F. Barbaresco, and L. Yang, “Medians and Means in Riemannian Geometry: Existence, Uniqueness and Computation,” in Matrix Information Geometry (F. Nielsen and R. Bhata, eds.), pp. 169–197, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.
[17] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, Model-Based Clustering and Classification for Data Science: With Applications in R. No. 50 in Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, United Kingdom New York, NY: Cambridge University Press, 2019.