

# ROBUST OPTIMIZATION AND INFERENCE ON MANIFOLDS

LIZHEN LIN, DREW LAZAR, BAYAN SARPABAYEVA, AND DAVID B. DUNSON

**ABSTRACT.** We propose a robust and scalable procedure for general optimization and inference problems on manifolds leveraging the classical idea of ‘median-of-means’ estimation. This is motivated by ubiquitous examples and applications in modern data science in which a statistical learning problem can be cast as an optimization problem over manifolds. Being able to incorporate the underlying geometry for inference while addressing the need for robustness and scalability presents great challenges. We address these challenges by first proving a key lemma that characterizes some crucial properties of geometric medians on manifolds. In turn, this allows us to prove robustness and tighter concentration of our proposed final estimator in a subsequent theorem. This estimator aggregates a collection of subset estimators by taking their geometric median over the manifold. We illustrate bounds on this estimator via calculations in explicit examples. The robustness and scalability of the procedure is illustrated in numerical examples on both simulated and real data sets.

**KEYWORDS:** Geometric median on manifolds; Median-of-means; Optimization on manifolds; Robust inference; Robust principal geodesic analysis (RPGA); Scalability

## 1. INTRODUCTION

There is a rapidly growing collection of learning problems and applications in data science that can be formalized as optimization problems over non-Euclidean spaces, such as non-linear Riemannian manifolds. Advancement in technology and computing leads to the increasing prevalence of complex data that are in non-Euclidean forms, such as positive definite matrices (diffusion matrices) in diffusion tensor imaging [1], shape objects in medical vision [16], network data objects [17], subspaces or orthonormal frames and so on [21]. Proper statistical inference from such data involves optimization over the underlying manifold to which the data are constrained. For example, there is a vibrant line of research based on estimation of Fréchet means [12], which are minimizers of Fréchet functions on manifolds [5, 6]. In this case, both the data and parameters of interest are on manifolds. In addition, it is common to represent lower-dimensional structure in high-dimensional data as a manifold. Learning such a manifold is a non-trivial optimization problem. In each of the above problems, developing algorithms that are robust to data contamination and heavy tails and that scale efficiently to large datasets is crucial.

With this motivation, our main aim is to propose a robust and scalable procedure for general optimization on manifolds. We generalize the powerful ‘median-of-means’ estimator [29], to manifolds by establishing some key properties of the geometric median on manifolds with which we can prove tighter concentration bounds of our proposed estimator. The key idea is to obtain optimizers from subset data which are aggregated to form a final estimator. Our estimator can be shown to be robust to outliers and contaminations of arbitrary nature and has provable robustness. Scalability of the algorithm is automatically gained via the divide-and-conquer nature of combining subset-based estimators.

There is a related literature outside of the non-Euclidean manifold setting. For example, [26] applies the ‘median-of-means’ procedure for robust estimation in Banach spaces. In [27], a robust Bayesian estimator is proposed as the geometric median of subset posteriors measures. There has been recent theoretical and computational developments on applying the median of mean procedure in learning theory [19, 24]. Characterizing properties of the geometric median on manifolds requires a substantially different approach, which deals with the underlying geometry. We prove a key lemma characterizing the robustness property of geometric medians on manifolds, which allows us to show our estimator has tighter concentration bounds than subset estimators. This is done for both the *extrinsic geometric median* and the *intrinsic geometric median* with the former employing an embedding of manifolds into some higher-dimensional Euclidean space and the later adopting a Riemannian structure. We illustrate the bounds with explicit calculations in both the extrinsic and intrinsic cases. Our procedure is demonstrated in a class of manifolds through both simulated and real data examples. Manifolds considered include the sphere, positive definite matrices and the planar shape spaces, all of which are commonly applicable in real data analyses.

The paper is organized as follows: in section 2 we introduce the general procedure and prove a key property of the geometric median on manifolds. Section 3 is devoted to robust estimation and optimization on manifolds. In particular, we prove the concentration property of our final estimator in estimating the population parameter of interest and provide examples of calculations of the bounds. In section

4 we consider an extensive simulation study and data analysis illustrating both the robustness and scalability of our procedure. The paper ends with a discussion.

## 2. GEOMETRIC MEDIAN AND ROBUST ESTIMATION ON MANIFOLDS

Let  $Q$  be a probability distribution on some space  $\mathcal{X}$  and  $\mathcal{M}$  be a manifold. We consider the problem of estimating the *population parameter*

$$(2.1) \quad \mu = \arg \min_{p \in \mathcal{M}} L^*(p),$$

where  $L^*(p)$  is defined as

$$L^*(p) = \int_{\mathcal{X}} L(p, x) Q(dx)$$

for some loss function  $L$ .

Let  $\mathbf{x} = \{x_1, \dots, x_n\}$  where  $x_1, \dots, x_n$  are sampled from  $Q$ . The parameter  $\mu$  is often estimated by the *empirical risk estimator*

$$(2.2) \quad \hat{\mu}_n = \arg \min_{p \in \mathcal{M}} L_n(p, \mathbf{x}) = \arg \min_{p \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n L(p, x_i).$$

*Remark 2.1.* An important example is the *Fréchet mean* in which the risk function is

$$L^*(p) = \int \rho^2(p, x) Q(dx),$$

with  $Q$  supported on a manifold  $\mathcal{X} = \mathcal{M}$  and  $\rho$  a metric defined on  $\mathcal{M}$ , and  $\hat{\mu}_n$  corresponds to the sample Fréchet mean. There is significant literature on nonparametric statistical inference on manifolds in which estimation of the Fréchet mean is addressed (see [5, 6]). Similarly, in a regression problem with manifold-valued output, the underlying problem can be cast as an optimization problem on manifolds [22]. In many other applications, we do not have  $\mathcal{X} = \mathcal{M}$  with  $\mathcal{X}$  a higher-dimensional ambient space and optimization done over a lower-dimensional manifold such as the Grassmannian [23, 31], which has abundant applications in manifold learning and low-rank estimation matrix problems [7, 9].

Real data sets often contain outliers that can be errors, extreme observations or contamination of various sorts which occur when sampling from heavy tailed or mixture distributions. Thus, there is interest in robust estimation of population parameters by estimators which are stable and not unduly effected by the presence of outliers.

In this paper, we consider the classic and intuitive estimator formed by taking the geometric median of a collection of subset estimators or optimizers. Before formally introducing our procedure in the next section, we introduce the notion of the *geometric median on a manifold* and prove an important lemma about its properties.

For a metric space  $(\mathcal{M}, \rho)$  the *geometric median*,  $p^*$ , of points  $p_1, \dots, p_m \in \mathcal{M}$  minimizes the sum of distances to the points, i.e.,

$$(2.3) \quad p^* = \text{med}(p_1, \dots, p_m) = \arg \min_{p \in \mathcal{M}} \frac{1}{m} \sum_{k=1}^m \rho(p, p_k)$$

assuming that  $p^*$  exists and is unique. When  $\mathcal{M}$  is a manifold, there are different ways to metrize the space. Let  $J : \mathcal{M} \rightarrow \mathbb{R}^D$  be an embedding of a manifold  $\mathcal{M}$  into

some higher-dimensional Euclidean space  $\mathbb{R}^D$ . One can define an *extrinsic distance* on  $M$  induced from the embedding  $J$  in which

$$\rho(p, q) = \|J(p) - J(q)\|,$$

where  $\|\cdot\|$  is the Euclidean norm on  $\mathbb{R}^D$ . Alternatively, one can take  $\rho$  to be the *intrinsic distance* as the geodesic distance arising from a Riemannian structure on  $\mathcal{M}$ .

With the choice of  $\rho$  as the extrinsic or intrinsic distance in (2.3), we have corresponding definitions of the *extrinsic geometric median* and the *intrinsic geometric median*, respectively. Some properties of the intrinsic geometric median are studied in [11] by, for example, characterizing the uniqueness conditions of the intrinsic sample median along with a Weizfeld algorithm for finding the median. Our theoretical results below on robustness are of a fundamentally different nature, allowing us to construct an estimator that is not only robust but also has tighter bounds around the true parameter of interest.

We prove the following lemma, which says if  $\omega \in \mathcal{M}$  is at least a constant,  $C_\alpha$ , times  $\epsilon$  distance away from the geometric median  $p^* = \text{med}(p_1, \dots, p_m)$ , then  $\omega$  is at least  $\epsilon$  distance away from at least an  $\alpha$  fraction of the points  $p_1, \dots, p_m$ . This result is illustrated in Figure 1. A similar result was proved in [26] in the case of Banach spaces. The proof of the following, a general lemma for manifolds, requires additional machinery.

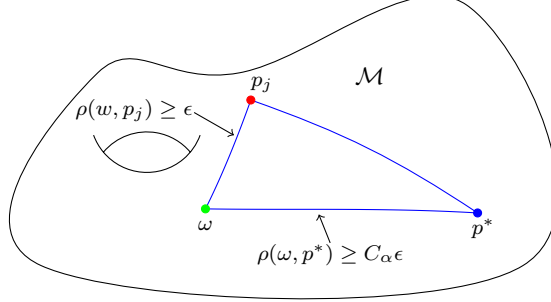


FIGURE 1. Geometric Illustration of Lemma 2.1 on Manifold  $\mathcal{M}$

**Lemma 2.1.** *Let  $p_1, \dots, p_m \in \mathcal{M}$ ,  $p^* = \text{med}(p_1, \dots, p_m)$  as in (2.3). Then (a) and (b) below hold.*

- (a) *Let  $\rho$  be the extrinsic distance for some embedding  $J : \mathcal{M} \rightarrow \tilde{\mathcal{M}} \subset \mathbb{R}^D$ . Let  $\omega \in \mathcal{M}$ ,  $\psi$  be angle between  $J(\omega) - J(p^*)$  and the tangent space  $T_{J(p^*)}\tilde{\mathcal{M}}$  and let*

$$C_\alpha = \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \psi} - \alpha \sin \psi}$$

*where  $\alpha \in (0, \cot \psi \tan \frac{\psi}{2})$ . If  $\rho(\omega, p^*) \geq C_\alpha \epsilon$ , then there exists an  $\alpha$  portion of elements of  $p_1, \dots, p_m$  which are at least  $\epsilon$  distance away from  $\omega$ . That is, there exists an index set  $T \subset \{1, \dots, m\}$  with  $|T| \geq \alpha m$ , and  $\rho(p_j, \omega) \geq \epsilon$  for any  $j \in T$ .*

- (b) Let  $\rho$  be an intrinsic distance on  $M$  with respect to some Riemannian structure. Let  $\omega \in \mathcal{M}$ , the log map  $\log_{p^*}$  be  $K$ -Lipschitz continuous from  $B(\omega, \epsilon)$  to  $T_{p^*}\mathcal{M}$  and let

$$C_\alpha = K(1 - \alpha)\sqrt{\frac{1}{1 - 2\alpha}}$$

where  $\alpha \in (0, 1/2)$ . If  $\rho(\omega, p^*) \geq C_\alpha \epsilon$ , then there exists an  $\alpha$  portion of elements of  $p_1, \dots, p_m$  which are at least  $\epsilon$  distance away from  $\omega$ .

*Proof.* (a) Let  $L(J(p)) = \sum_{j=1}^m \rho(p, p_j) = \sum_{j=1}^m \|J(p) - J(p_j)\|$  for  $J(p) \in \tilde{\mathcal{M}}$ . Let  $\gamma(t)$  be a curve from  $J(p^*)$  to  $J(\omega)$  on  $\tilde{\mathcal{M}}$ , where  $\gamma(0) = J(p^*)$ ,  $\gamma(1) = J(\omega)$ , and  $\gamma'(0) = v$ . The directional derivative of  $L$  at  $J(p^*)$  evaluated at  $v$  is given by

$$\begin{aligned} dL_{J(p^*)}(v) &= \lim_{t \rightarrow 0^+} \frac{L(\gamma(t)) - L(\gamma(0))}{t} \\ (2.4) \quad &= \lim_{t \rightarrow 0^+} \frac{L(\gamma(t)) - L(J(p^*))}{t} \geq 0 \end{aligned}$$

with the above inequality holding as  $J(p^*)$  minimizes  $L$  for  $p \in \mathcal{M}$ . Let

$$\gamma(t) = \mathcal{P}_{\tilde{\mathcal{M}}}\left(J(p^*) + t(J(\omega) - J(p^*))\right),$$

where  $\mathcal{P}$  is the projection of  $\mathbb{R}^D$  onto  $\tilde{\mathcal{M}}$ , that is,

$$\mathcal{P}(x) = \arg \min_{y \in \tilde{\mathcal{M}}} \rho(y, x).$$

We assume the projection map  $\mathcal{P}$  is differentiable at  $t = 0$ . Denote  $\mathcal{J}$  as the Jacobian matrix of the projection map  $\mathcal{P}$  at  $J(p^*)$ . Then one has

$$v = \gamma'(0) = \mathcal{J}(J(\omega) - J(p^*)),$$

which will be needed in determining the constant  $C_\alpha$ . One can see that

$$L(\gamma(t)) - L(J(p^*)) = \sum_{j=1}^m (\|\gamma(t) - J(p_j)\| - \|\gamma(0) - J(p_j)\|).$$

Let

$$A_j = \frac{\|\gamma(t) - J(p_j)\| - \|\gamma(0) - J(p_j)\|}{t} \text{ for } j = 1, \dots, m.$$

Then

$$A_j = \frac{\|\gamma(t) - J(p_j)\|^2 - \|\gamma(0) - J(p_j)\|^2}{t(\|\gamma(t) - J(p_j)\| + \|\gamma(0) - J(p_j)\|)} \text{ for } j = 1, \dots, m.$$

One has

$$(2.5) \quad \lim_{t \rightarrow 0^+} (\|\gamma(t) - J(p_j)\| + \|\gamma(0) - J(p_j)\|) = 2\|\gamma(0) - J(p_j)\|.$$

Also,

$$\begin{aligned} \|\gamma(t) - J(p_j)\|^2 &= \langle \gamma(t) - J(p_j), \gamma(t) - J(p_j) \rangle \\ &= \langle \gamma(t), \gamma(t) \rangle - 2\langle \gamma(t), J(p_j) \rangle + \langle J(p_j), J(p_j) \rangle, \end{aligned}$$

and

$$\|\gamma(0) - J(p_j)\|^2 = \langle \gamma(0), \gamma(0) \rangle - 2\langle \gamma(0), J(p_j) \rangle + \langle J(p_j), J(p_j) \rangle.$$

Then

$$\begin{aligned}
& \|\gamma(t) - J(p_j)\|^2 - \|\gamma(0) - J(p_j)\|^2 \\
&= \langle \gamma(t), \gamma(t) \rangle - \langle \gamma(0), \gamma(0) \rangle - 2\langle \gamma(t) - \gamma(0), J(p_j) \rangle \\
&= (\langle \gamma(t), \gamma(t) \rangle - \langle \gamma(0), \gamma(t) \rangle) + (\langle \gamma(0), \gamma(t) \rangle - \langle \gamma(0), \gamma(0) \rangle) \\
&\quad - 2\langle \gamma(t) - \gamma(0), J(p_j) \rangle \\
&= \langle \gamma(t) - \gamma(0), \gamma(t) \rangle + \langle \gamma(0), \gamma(t) - \gamma(0) \rangle - 2\langle \gamma(t) - \gamma(0), J(p_j) \rangle \\
&= \langle \gamma(t) - \gamma(0), \gamma(t) + \gamma(0) - 2J(p_j) \rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{t \rightarrow 0^+} \frac{\|\gamma(t) - J(p_j)\|^2 - \|\gamma(0) - J(p_j)\|^2}{t} &= \lim_{t \rightarrow 0^+} \left\langle \frac{\gamma(t) - \gamma(0)}{t}, \gamma(t) + \gamma(0) - 2J(p_j) \right\rangle \\
&= \langle \gamma'(0), \gamma(0) + \gamma(0) - 2J(p_j) \rangle \\
&= 2\langle \gamma'(0), \gamma(0) - J(p_j) \rangle = 2\langle \gamma'(0), J(p^*) - J(p_j) \rangle.
\end{aligned}$$

Thus, by (2.5) and the above equation, if  $J(p_j) \neq J(p^*)$ , one has

$$\lim_{t \rightarrow 0^+} A_j = \frac{\langle \gamma'(0), J(p^*) - J(p_j) \rangle}{\|J(p^*) - J(p_j)\|}.$$

Otherwise, if  $J(p_j) = J(p^*)$ , then

$$\lim_{t \rightarrow 0^+} A_j = \lim_{t \rightarrow 0^+} \frac{\|\gamma(t) - J(p_j)\|}{t} = \|\gamma'(0)\|.$$

Therefore,

$$\begin{aligned}
dL_{J(p^*)}(v) &= \sum_{j=1}^m \lim_{t \rightarrow 0^+} A_j \\
&= \sum_{j: p_j \neq p^*} \frac{\langle \gamma'(0), J(p^*) - J(p_j) \rangle}{\|J(p^*) - J(p_j)\|} + \|\gamma'(0)\| \sum_{j=1}^m I(p_j = p^*),
\end{aligned}$$

where  $I(\cdot)$  is the indicator function. The above implies

$$(2.6) \quad \frac{dL_{p^*}(v)}{\|\gamma'(0)\|} = \sum_{j=1}^m \lim_{t \rightarrow 0^+} \frac{A_j}{\|\gamma'(0)\|}$$

$$(2.7) \quad = \sum_{j: p_j \neq p^*} \frac{\langle \gamma'(0), J(p^*) - J(p_j) \rangle}{\|\gamma'(0)\| \|J(p^*) - J(p_j)\|} + \sum_{j=1}^m I(p_j = p^*).$$

The Jacobian matrix of the projection map  $\mathcal{P}$  at  $J(p^*)$ ,  $\mathcal{J}$ , is the orthogonal projection of  $T_{J(p^*)}\mathbb{R}^D \equiv \mathbb{R}^D$  to  $T_{J(p^*)}\tilde{\mathcal{M}}$ . That is, for  $a \in T_{J(p^*)}\mathbb{R}^D$ ,  $\mathcal{J}(a) = a_1$ , where  $a = a_1 + a_2$  is the unique orthogonal decomposition of  $a$  with  $a_1 \in T_{J(p^*)}\tilde{\mathcal{M}}$ . Now assume that there does *not* exist an  $\alpha$  portion of elements of  $p_1, \dots, p_m$  which are at least  $\epsilon$  distance away from  $\omega$ , that is, without loss of generality,

$$\|J(p_j) - J(\omega)\| \leq \epsilon \text{ for } j = 1, \dots, \lfloor (1 - \alpha)m \rfloor + 1.$$

Let us denote by  $\angle(J(\omega) - J(p^*), J(p_j) - J(p^*))$  the angle between the vectors  $J(\omega) - J(p^*)$  and  $J(p_j) - J(p^*)$ . Then for  $j = 1, \dots, \lfloor (1 - \alpha)m \rfloor + 1$ ,

$$\sin \left( \angle(J(\omega) - J(p^*), J(p_j) - J(p^*)) \right) < \frac{1}{C_\alpha}$$

and so

$$\cos \left( \angle(J(\omega) - J(p^*), J(p_j) - J(p^*)) \right) > \sqrt{1 - \frac{1}{C_\alpha^2}}.$$

Notice that

$$\begin{aligned} & \angle \left( \mathcal{J}(J(\omega)) - J(p^*), J(\omega) - J(p^*) \right) + \angle(J(\omega) - J(p^*), J(p_j) - J(p^*)) \\ &= \psi + \angle(J(\omega) - J(p^*), J(p_j) - J(p^*)) \geq \angle \left( \mathcal{J}(J(\omega)) - J(p^*), J(p_j) - J(p^*) \right). \end{aligned}$$

Therefore,

$$\begin{aligned} \cos \left( \angle \left( \mathcal{J}(J(\omega)) - J(p^*), J(p_j) - J(p^*) \right) \right) &\geq \cos \left( \psi + \angle(J(\omega) - J(p^*), J(p_j) - J(p^*)) \right) \\ &> \sqrt{1 - \frac{1}{C_\alpha^2}} \cos \psi - \frac{1}{C_\alpha} \sin \psi. \end{aligned}$$

We have

$$\begin{aligned} \frac{\langle \gamma'(0), p_j - J(p^*) \rangle}{\|\gamma'(0)\| \|p_j - J(p^*)\|} &= \cos \left( \angle \left( \mathcal{J}(J(\omega)) - J(p^*), J(p_j) - J(p^*) \right) \right) \\ &> \sqrt{1 - \frac{1}{C_\alpha^2}} \cos \psi - \frac{1}{C_\alpha} \sin \psi. \end{aligned}$$

Then for any  $\alpha \in \left(0, \cot \psi \tan \frac{\psi}{2}\right)$  from (2.6)

$$\frac{dL_{J(p^*)}(v)}{\|\gamma'(0)\|} < -(1 - \alpha)m \left( \sqrt{1 - \frac{1}{C_\alpha^2}} \cos \psi - \frac{1}{C_\alpha} \sin \psi \right) + \alpha m \leq 0,$$

when

$$C_\alpha \geq \frac{1 - \alpha}{\sqrt{1 - 2\alpha} \cos \psi - \alpha \sin \psi}$$

which is a contradiction with (2.4).

(b) The intrinsic median requires a different proof. Let  $L(p) = \sum_{j=1}^m \rho(p, p_j)$  where  $\rho$  is the intrinsic distance; we use the Riemannian exponential map  $\exp_{p^*} : T_{p^*}\mathcal{M} \rightarrow \mathcal{M}$ . Let  $v = \log_{p^*} \omega \in T_{p^*}\mathcal{M}$  and consider the geodesic curve  $\gamma(t) = \exp_{p^*}(tv)$ . Then

$$(2.8) \quad dL_{p^*}(v) = \lim_{t \rightarrow 0} \frac{L(\gamma(t)) - L(\gamma(0))}{t} = \lim_{t \rightarrow 0} \frac{L(\gamma(t)) - L(p^*)}{t} \geq 0.$$

Denote

$$A = \lim_{t \rightarrow 0+} \sum_{j=1}^m \left( \frac{\sqrt{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle} - \sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}}{t} \right),$$

where  $\gamma_j(s, t) = \exp_{\gamma(t)}(s \log_{\gamma(t)} p_j) = \exp_{\gamma(t)}(s v_j(t))$  is the geodesic curve connecting  $\gamma(t)$  with  $p_j$ , then  $\gamma_{js}(s, t) = \frac{\partial \gamma_j(s, t)}{\partial s}$ . Set

$$A_j = \frac{\sqrt{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle} - \sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}}{t}, \text{ for } j = 1, \dots, m.$$

Then

$$A_j = \frac{1}{t} \frac{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle - \langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}{\sqrt{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle} + \sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}}, \text{ for } j = 1, \dots, m.$$

We see that

$$\lim_{t \rightarrow 0^+} \left( \sqrt{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle} + \sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle} \right) = 2\sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}.$$

On the other hand,

$$\begin{aligned} \lim_{t \rightarrow 0^+} \frac{\langle \gamma_{js}(s, t), \gamma_{js}(s, t) \rangle - \langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}{t} &= 2 \left\langle \frac{D}{dt} \gamma_{js}(s, 0), \gamma_{js}(s, 0) \right\rangle \\ &= 2 \left\langle \frac{D}{ds} \gamma_{jt}(s, 0), \gamma_{js}(s, 0) \right\rangle = 2 \frac{d}{ds} \langle \gamma_{jt}(s, 0), \gamma_{js}(s, 0) \rangle. \end{aligned}$$

Thus if  $p_j \neq p^*$ , one has

$$\lim_{t \rightarrow 0^+} A_j = \frac{\frac{d}{ds} \langle \gamma_{jt}(s, 0), \gamma_{js}(s, 0) \rangle}{\sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}}.$$

Otherwise, if  $p_j = p^*$ , then

$$\lim_{t \rightarrow 0^+} A_j = \lim_{t \rightarrow 0^+} \frac{\sqrt{\langle -t\gamma'((1-s)t), -t\gamma'((1-s)t) \rangle}}{t} = \lim_{t \rightarrow 0^+} \frac{t\|v\|}{t} = \|v\|.$$

Therefore,

$$\begin{aligned} dL_{p^*}(v) &= \sum_{j=1}^m \int_0^1 \lim_{t \rightarrow 0^+} A_j ds \\ &= \sum_{j: p_j \neq p^*} \int_0^1 \frac{\frac{d}{ds} \langle \gamma_{jt}(s, 0), \gamma_{js}(s, 0) \rangle}{\sqrt{\langle \gamma_{js}(s, 0), \gamma_{js}(s, 0) \rangle}} ds + \|v\| \sum_{j=1}^m I(p_j = p^*) \\ &= \sum_{j: p_j \neq p^*} \frac{\langle \gamma_{jt}(1, 0), \gamma_{js}(1, 0) \rangle}{\|v_j\|} + \|v\| \sum_{j=1}^m I(p_j = p^*) \\ &= \sum_{j: p_j \neq p^*} \frac{\langle (d \exp_{p^*})_{v_j}(1 \cdot v'_j(0)), (d \exp_{p^*})_{v_j} v_j \rangle}{\|v_j\|} + \|v\| \sum_{j=1}^m I(p_j = p^*) \\ &= \sum_{j: p_j \neq p^*} \frac{\langle v'_j(0), v_j \rangle}{\|v_j\|} + \|v\| \sum_{j=1}^m I(p_j = p^*) \\ &= - \sum_{j: p_j \neq p^*} \frac{\langle v, v_j \rangle}{\|v_j\|} + \|v\| \sum_{j=1}^m I(p_j = p^*), \end{aligned}$$

where  $I(\cdot)$  is the indicator function. Then one has,

$$\begin{aligned} \frac{dL_{p^*}(v)}{\|v\|} &= - \sum_{j: p_j \neq p^*} \frac{\langle v, v_j \rangle}{\|v\| \|v_j\|} + \sum_{j=1}^m I(p_j = p^*) \\ &= - \sum_{j: p_j \neq p^*} \cos(\widehat{v, v_j}) + \sum_{j=1}^m I(p_j = p^*). \end{aligned}$$

From the condition that  $\log_{p^*}$  is  $K$ -Lipschitz continuous from  $B(\omega, r)$  to  $T_{p^*}\mathcal{M}$ ,

$$\|v_j - v\| \leq K d_g(\exp_{p^*} v_j, \exp_{p^*} v).$$

Then this yields

$$\frac{dL_{p^*}(v)}{\|v\|} < -(1 - \alpha)m\sqrt{1 - \frac{K^2}{C_\alpha^2}} + \alpha m \leq 0,$$

whenever  $C_\alpha \geq K(1 - \alpha)\sqrt{\frac{1}{1 - 2\alpha}}$ , which leads to a contradiction with (2.8).  $\square$

There are many known Riemannian manifolds with  $K$ -Lipschitz continuous log maps as required in part (b) of the above lemma. Below we provide a few examples including the sphere, the planar shape space and the space of positive definite matrices, which are commonly encountered manifolds in the statistics and medical imaging literature.

**Proposition 2.1.** *Let  $S^d = \{p \in \mathbb{R}^{d+1} : \|p\| = 1\}$  which is the  $d$ -dimensional sphere. The inverse exponential map,  $\log_p$ , on  $S^d$  is 2-Lipschitz continuous from  $B(p, \pi/2)$  to  $T_p S^d$  for all  $p \in S^d$ .*

*Proof.* The tangent space at  $p$  is given as

$$T_p S^d = \{v \in \mathbb{R}^{d+1} : v^T p = 0\}.$$

Then for  $q \in S^d$  the inverse exponential map can be expressed as

$$\log_p(q) = \frac{\arccos(p^T q)}{\sqrt{1 - (p^T q)^2}}(q - (p^T q)p).$$

Hence, the distance between  $\log_p q_1$  and  $\log_p q_2$  is equal to

$$\|\log_p q_1 - \log_p q_2\| = \sqrt{\arccos(p^T q_1)^2 + \arccos(p^T q_2)^2 - 2 \arccos(p^T q_1) \arccos(p^T q_2) \cos \varphi}$$

where  $\varphi$  is the angle between  $\log_p q_1$  and  $\log_p q_2$ . The geodesic distance between  $q_1$  and  $q_2$  is then given by

$$d_g(q_1, q_2) = \arccos(q_1^T q_2).$$

One can easily obtain that

$$q_1^T q_2 = (p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi.$$

Then one can check directly that

$$\|\log_p q_1 - \log_p q_2\| \leq 2d_g(q_1, q_2).$$

$\square$

The following proposition shows that the log map in similarity-shape spaces [16] also satisfies the  $K$ -Lipschitz condition.

**Proposition 2.2.** *The similarity or planar shape space is given as*

$$(2.9) \quad \Sigma_2^k = S^{2k-3}/S^1.$$

*The inverse exponential map,  $\log_p$ , on  $\Sigma_2^k$  is 2-Lipschitz continuous from  $B(p, \pi/4)$  to  $T_p \Sigma_2^k$  for all  $p \in \Sigma_2^k$ .*

*Proof.*  $\Sigma_2^k$  is the quotient of the sphere  $S^{2k-3}$  under the following group of transformations

$$G = \left\{ \begin{pmatrix} A & \dots & 0 \\ & \ddots & \\ 0 & \dots & A \end{pmatrix} \in M(2k), \quad A \in SO(2) \right\} \simeq S^1.$$

For any  $B \in G$ , we have that  $B = \cos tI + \sin t\dot{I}$ , where

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \quad \dot{I} = \begin{pmatrix} 0 & 1 & \dots & 0 & 0 \\ -1 & 0 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & -1 & 0 \end{pmatrix}.$$

For each  $p \in \Sigma_2^k$  we define the tangent space

$$T_p \Sigma_2^k = \{v \in \mathbb{R}^{2k-2} : v^T p = 0, (Ip)^T v = 0\}.$$

The inverse exponential map can be expressed as

$$\log_p(q) = \frac{\arccos(p^T q)}{\sqrt{1 - (p^T q)^2}} (q - (p^T q)p).$$

Hence, the distance between  $\log_p q_1$  and  $\log_p q_2$  is equal to

$$\|\log_p q_1 - \log_p q_2\| = \sqrt{\arccos(p^T q_1)^2 + \arccos(p^T q_2)^2 - 2 \arccos(p^T q_1) \arccos(p^T q_2) \cos \varphi},$$

where  $\varphi$  is an angle between  $\log_p q_1$  and  $\log_p q_2$ . The geodesic distance between  $q_1$  and  $q_2$  is then given by

$$\begin{aligned} d_g(q_1, q_2) &= \inf_{t \in (-\pi, \pi]} \arccos(q_1^T (\cos tI + \sin t\dot{I})q_2) \\ &= \arccos \sup_{t \in (-\pi, \pi]} (\cos tq_1^T q_2 + \sin tq_1^T \dot{I}q_2) \\ &= \arccos \sqrt{(q_1^T q_2)^2 + (q_1^T \dot{I}q_2)^2}. \end{aligned}$$

One can easily obtain that

$$\begin{aligned} (q_1^T q_2)^2 + (q_1^T \dot{I}q_2)^2 &= ((p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi)^2 \\ &\quad + (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\cos \psi)^2 \end{aligned}$$

where  $\psi$  is the angle between  $\log_p q_1$  and  $\dot{I} \log_p q_2$ . Note that

$$\pi/2 - \varphi \leq \psi \leq \pi/2 + \varphi.$$

Thus  $\cos \psi \geq \cos(\pi/2 - \varphi) = \sin \varphi$ , and

$$\begin{aligned} d_g(q_1, q_2) &\geq 2 \arccos \left( ((p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi)^2 \right. \\ &\quad \left. + (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\sin \varphi)^2 \right)^{1/2}. \end{aligned}$$

Then it can be verified directly that

$$\begin{aligned} \|\log_p q_1 - \log_p q_2\| &\leq 2 \arccos \left( ((p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi)^2 \right. \\ &\quad \left. + (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\sin \varphi)^2 \right)^{1/2}. \end{aligned}$$

Thus  $\|\log_p q_1 - \log_p q_2\| \leq 2d_g(q_1, q_2)$ .  $\square$

**Proposition 2.3.** *The manifold of positive definite  $n$  by  $n$  matrices  $\text{PD}(n)$  has a 1-Lipchitz continuous inverse exponential map at any  $p \in \text{PD}(n)$ .*

*Proof.* We consider the Killing metric [30] in the manifold of invertible  $n$  by  $n$  matrices  $\text{GL}(n)$

$$ds^2(a) = \text{tr}(a^{-1}da)^2.$$

In other words, in the Lie algebra  $\mathfrak{gl}(n) = T_I \text{GL}(n) = \text{M}(n)$ , we have the symmetric inner product

$$\langle A, B \rangle_I = \text{tr}(AB), \quad A, B \in \mathfrak{gl}(n),$$

which generates the bilaterally invariant metric in the group  $\text{GL}(n)$ . That is, for any  $A, B \in T_g \text{GL}(n)$  and  $a \in \text{GL}(n)$ ,

$$\langle A, B \rangle_a = \langle a^{-1}A, a^{-1}B \rangle_I = \langle Aa^{-1}, Ba^{-1} \rangle_I = \text{tr}(a^{-1}Aa^{-1}B).$$

Since vectors  $p^{-1}A, p^{-1}B$  do not always belong to the tangent space  $T_I \text{PD}(n)$ , we instead take vectors  $p^{-1/2}Ap^{-1/2}$  and

$$\begin{aligned} \langle A, B \rangle_p &= \langle p^{-1/2}A, p^{-1/2}B \rangle_{p^{1/2}} \\ &= \langle p^{-1/2}Ap^{-1/2}, p^{-1/2}Bp^{-1/2} \rangle_I \\ &= \text{tr}(p^{-1/2}Ap^{-1}Bp^{-1/2}) = \text{tr}(p^{-1}Ap^{-1}B), \end{aligned}$$

where  $A, B \in T_p \text{PD}(n)$ . Hence we have the metric in  $\text{PD}(n)$  induced from the Killing metric in  $\text{GL}(n)$ . This metric is usually known as the Fisher-Rao metric.

For this metric we have the following exponential and logarithm mappings

$$\begin{aligned} \exp_p A &= p^{1/2} \exp(p^{-1/2}Ap^{-1/2})p^{1/2}, \\ \log_p q &= p^{1/2} \log(p^{-1/2}qp^{-1/2})p^{1/2}, \end{aligned}$$

where

$$\begin{aligned} \exp Y &= I + \frac{Y}{1!} + \frac{Y^2}{2!} + \dots + \frac{Y^n}{n!} + \dots, \\ \log x &= (x - I) - \frac{(x - I)^2}{2} + \dots + (-1)^{n-1} \frac{(x - I)^n}{n} + \dots \end{aligned}$$

for any  $A, Y \in \text{Sym}(n)$  and  $p, q, x \in \text{PD}(n)$ .

Let  $a, q_1, q_2 \in \text{PD}(n)$ . Then we have

$$\begin{aligned} \|\log_a q_1 - \log_a q_2\|_a &= \|\log(a^{-1/2}q_1a^{-1/2}) - \log(a^{-1/2}q_2a^{-1/2})\|_I \\ &\leq d_g(a^{-1/2}q_1a^{-1/2}, a^{-1/2}q_2a^{-1/2}) = d_g(q_1, q_2) \end{aligned}$$

where the inequality follows from the exponential metric increasing property of the Fisher-Rao metric as in [4].  $\square$

### 3. ROBUST OPTIMIZATION ON MANIFOLDS: CONCENTRATION PROPERTIES

In this section, we introduce our proposed estimator, which aggregates a collection of subset optimizers of the empirical risk function. We first divide the data set  $x_1, \dots, x_n$  into  $m$  subsets  $U_1, \dots, U_m$  each of roughly size  $\lfloor n/m \rfloor$ . Let  $\mu_1, \dots, \mu_m$  be the optimizers of the empirical risk function from each subset,  $U_1, \dots, U_m$ , respectively. That is,

$$(3.1) \quad \mu_j = \arg \min_{p \in \mathcal{M}} L_{|U_j|}(p, U_j) \text{ for } j = 1, \dots, m$$

as in (2.2). Our estimator  $\mu^*$  is the *geometric median of the subset optimizers*, that is,

$$(3.2) \quad \mu^* = \arg \min_{p \in \mathcal{M}} \sum_{j=1}^m \rho(p, \mu_j).$$

We will show that  $\mu^*$  has desired robustness properties in estimating the population parameter  $\mu$ .

In [26] it is proven that the geometric median of a collection of weakly concentrated estimators admits a tighter deviation bound in a Hilbert space. With the help of the Lemma 1, we generalise this result to manifolds in the following theorem.

**Theorem 3.1.** *Let  $\mu_1, \dots, \mu_m$  be a collection of independent estimators of the parameter  $\mu$ , and let geometric median  $\mu^* = \text{med}(\mu_1, \dots, \mu_m)$ .*

- (a) *Let  $\rho$  be the extrinsic distance on  $\mathcal{M}$  for some embedding  $J : \mathcal{M} \rightarrow \tilde{\mathcal{M}} \subset \mathbb{R}^D$ . Assume for any  $\omega \in \mathcal{M}$  the angle between  $J(\omega) - J(\mu^*)$  and the tangent space  $T_{J(\mu^*)}\tilde{\mathcal{M}}$  is no bigger than  $\bar{\psi}$ . For any  $\alpha \in (0, \cot \bar{\psi} \tan \frac{\bar{\psi}}{2})$  set*

$$\bar{C}_\alpha = \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \bar{\psi}} - \alpha \sin \bar{\psi}}.$$

- (b) *Let  $\rho$  be an intrinsic distance on  $\mathcal{M}$  with respect to some Riemannian structure. Assume  $\log_{\mu^*}$  is  $K$ -Lipschitz continuous from  $B(\mu^*, \epsilon)$  to  $T_{\mu^*}\mathcal{M}$ . For any  $\alpha \in (0, \frac{1}{2})$  set*

$$\bar{C}_\alpha = K(1 - \alpha) \sqrt{\frac{1}{1 - 2\alpha}}.$$

Under (a) or (b), if

$$(3.3) \quad P(\rho(\mu_j, \mu) > \epsilon) \leq \eta \text{ for } i = 1, \dots, n$$

where  $\eta < \alpha$  then

$$(3.4) \quad P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)),$$

where

$$\phi(\alpha, \eta) = (1 - \alpha) \log \frac{1 - \alpha}{1 - \eta} + \alpha \log \frac{\alpha}{\eta}.$$

*Proof.* Let  $\psi$  be the angle between  $J(\mu) - J(\mu^*)$  and the tangent space  $T_{\mu^*}\tilde{\mathcal{M}}$ . Since  $\psi < \bar{\psi}$  we have  $C_\alpha \leq \bar{C}_\alpha$  and  $\cot \bar{\psi} \tan \frac{\bar{\psi}}{2} \leq \cot \psi \tan \frac{\psi}{2}$  where

$$C_\alpha = \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \psi} - \alpha \sin \psi}.$$

Thus, when the event  $\{\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon\}$  occurs, the event  $\{\rho(\mu^*, \mu) > C_\alpha \epsilon\}$  occurs. Then, by Lemma 1, when the event  $\{\rho(\mu^*, \mu) > C_\alpha \epsilon\}$  occurs, there exists

an  $\alpha$  portion of elements of  $\mu_1, \dots, \mu_m$  which are at least  $\epsilon$  distance away from  $\mu$ . Therefore,

$$(3.5) \quad P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq P(\rho(\mu^*, \mu) > C_\alpha \epsilon) \leq P\left(\sum_{j=1}^m I_{\rho(\mu_j, \mu) > \epsilon} > \alpha m\right).$$

Let  $A = |\{j = 1, \dots, m : \rho(\mu_j, \mu) > \epsilon\}|$  and let  $B$  be a random variable with a binomial distribution,  $B \sim b(m, \eta)$ . Then with (3.3) and by Lemma 23 in [20] there exists a coupling  $C = (\tilde{A}, \tilde{B})$  such that  $\tilde{A}$  has the same distribution as  $A$  and  $\tilde{B}$  has the same distribution as  $B$  such that  $\tilde{A} \leq \tilde{B}$ . Hence

$$P(A > \alpha m) \leq P(B > \alpha m) \leq \exp(-m\phi(\alpha, \eta))$$

where the second inequality follows from Chernoff's bound. Then with (3.5) we have

$$P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)).$$

For the intrinsic case (b) we have a similar proof.  $\square$

*Remark 3.1.* One important aspect in constructing the estimator  $\mu^*$  is the choice of the number of subsets  $m$ . By (3.4), a larger number of subset estimators,  $m$ , yields more robustness and a tighter concentration around the true parameter. At the same time, there must be enough data in each subset to ensure that each subset estimator behaves well and  $\eta$  in (3.3) is sufficiently small. For a given confidence level  $\epsilon$ , one can determine the number of subsets to achieve  $\eta$  in (3.3) and the desired bound on the concentration or confidence level in (3.4).

In the following, we provide examples, in both the intrinsic and extrinsic cases, of finding an  $\eta$  in (3.3) which allows the computation of the bound in (3.4).

*Example 1.* Consider the embedding  $J : \mathcal{M} \rightarrow \mathbb{R}^D$ . We have the induced measure  $\tilde{Q}$  on the image where  $\tilde{Q} = Q \circ J^{-1}$ . Let  $x_1, \dots, x_n$  be an i.i.d. sample from a distribution  $Q$ , such that we have the extrinsic mean  $\mu$  for the random variable  $x_1$

$$\mu = J^{-1}\left(\mathcal{P}\left(\int_{\mathbb{R}^D} u \tilde{Q}(du)\right)\right).$$

Divide the sample  $x_1, \dots, x_n$  into  $m$  disjoint groups  $U_1, \dots, U_m$  of size  $[n/m]$  each, and define

$$\begin{aligned} \tilde{\mu}_j &= \frac{1}{|U_j|} \sum_{i \in U_j} J(x_i) \quad j = 1, \dots, m, \\ \mu_j &\in J^{-1}(\mathcal{P}(\tilde{\mu}_j)). \end{aligned}$$

One can easily conclude that

$$\begin{aligned} \rho(\mu, \mu_j) &= \|J(\mu) - J(\mu_j)\| \\ &= \|J(\mu) - \tilde{\mu}_j + \tilde{\mu}_j - J(\mu_j)\| \\ &\leq \|J(\mu) - \tilde{\mu}_j\| + \|\tilde{\mu}_j - J(\mu_j)\| \\ &\leq 2\|J(\mu) - \tilde{\mu}_j\|. \end{aligned}$$

Therefore

$$\mathbb{E}\rho^2(\mu, \mu_j) \leq 4\mathbb{E}\|J(\mu) - \tilde{\mu}_j\|^2$$

$$\begin{aligned}
&= \frac{4}{|U_j|^2} \sum_{i \in U_j} \mathbb{E} \|J(\mu) - J(x_i)\|^2 \\
&\leq \frac{4}{|U_j|^2} \sum_{i \in U_j} \mathbb{E} \rho^2(\mu, x_i) \\
&= \frac{4}{|U_j|} \mathbb{E} \rho^2(\mu, x_1) \leq 4 \left\lceil \frac{m}{n} \right\rceil \mathbb{E} \rho^2(\mu, x_1).
\end{aligned}$$

So by Chebyshev's inequality

$$(3.6) \quad P(\rho(\mu_j, \mu) \geq \epsilon) = P(\rho^2(\mu_j, \mu) \geq \epsilon^2) \leq \frac{1}{\epsilon^2} \mathbb{E} \rho^2(\mu_j, \mu) \leq \frac{4}{\epsilon^2} \left\lceil \frac{m}{n} \right\rceil \mathbb{E} \rho^2(\mu, x_1).$$

Finally, we have the collection of independent estimators  $\mu_1, \dots, \mu_m$ , such that

$$P(\rho(\mu_j, \mu) > \epsilon) \leq \eta,$$

where  $\eta = \frac{4}{\epsilon^2} \left\lceil \frac{m}{n} \right\rceil \mathbb{E} \rho^2(\mu, x_1)$ . So by theorem 3.1 for any  $\alpha \in (0, \cot \bar{\psi} \tan \frac{\bar{\psi}}{2})$

$$P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)),$$

where

$$\begin{aligned}
\mu^* &= \text{med}(\mu_1, \dots, \mu_m), \\
\bar{C}_\alpha &= \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \bar{\psi} - \alpha \sin \bar{\psi}}}, \\
\phi(\alpha, \eta) &= (1 - \alpha) \log \frac{1 - \alpha}{1 - \eta} + \alpha \log \frac{\alpha}{\eta}.
\end{aligned}$$

*Example 2.* Let  $x_1, \dots, x_n$  be an i.i.d. sample from a distribution  $Q$ , such that we have the Fréchet mean  $\mu$  for the random variable  $x_1$ .

Divide the sample  $x_1, \dots, x_n$  into  $m$  disjoint groups  $U_1, \dots, U_m$  each of size  $\lceil n/m \rceil$ , and define

$$\mu_j = \arg \min_{y \in \mathcal{M}} \frac{1}{|U_j|} \sum_{i \in U_j} d_g^2(y, x_i), \quad j = 1, \dots, m.$$

Considering the  $j$ th subsample corresponding to  $U_j$  on the tangent space at  $\mu_j$ ,

$$\log_{\mu_j} \mu_j = \frac{1}{|U_j|} \sum_{x_i \in U_j} \log_{\mu_j} x_i = 0.$$

Thus on the tangent space  $T_{\mu_j} \mathcal{M}$ , we can obtain the equality

$$d_g^2(\mu, \mu_j) = \|\log_{\mu_j} \mu\|^2 = \frac{1}{|U_j|^2} \left\| \sum_{x_i \in U_j} (\log_{\mu_j} x_i - \log_{\mu_j} \mu) \right\|^2.$$

Thus,

$$\begin{aligned}
\mathbb{E} d_g^2(\mu, \mu_j) &= \frac{1}{|U_j|^2} \sum_{x_i \in U_j} \mathbb{E} \|\log_{\mu_j} x_i - \log_{\mu_j} \mu\|^2 \\
&\leq \frac{K^2}{|U_j|^2} \sum_{i \in U_j} \mathbb{E} d_g^2(\mu, x_i) = \frac{K^2}{|U_j|} \mathbb{E} d_g^2(\mu, x_1) \leq K^2 \left\lceil \frac{m}{n} \right\rceil \mathbb{E} d_g^2(\mu, x_1).
\end{aligned}$$

So by Chebyshev's inequality,  
(3.7)

$$P(d_g(\mu_j, \mu) \geq \epsilon) = P(d_g^2(\mu_j, \mu) \geq \epsilon^2) \leq \frac{1}{\epsilon^2} \mathbb{E} d_g^2(\mu_j, \mu) \leq \frac{K^2}{\epsilon^2} \left\lceil \frac{m}{n} \right\rceil \mathbb{E} d_g^2(\mu, x_1).$$

Finally, we have the collection of independent estimators  $\mu_1, \dots, \mu_m$ , such that

$$P(d_g(\mu_j, \mu) > \epsilon) \leq \eta,$$

where  $\eta = K^2 \left\lceil \frac{m}{n} \right\rceil \mathbb{E} d_g^2(\mu, x_1)$ . So by theorem 3.1 for any  $\alpha \in (0, \frac{1}{2})$

$$P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)),$$

where

$$\begin{aligned} \mu^* &= \text{med}(\mu_1, \dots, \mu_m), \\ \bar{C}_\alpha &= K(1 - \alpha) \sqrt{\frac{1}{1 - 2\alpha}}, \\ \phi(\alpha, \eta) &= (1 - \alpha) \log \frac{1 - \alpha}{1 - \eta} + \alpha \log \frac{\alpha}{\eta}. \end{aligned}$$

#### 4. SIMULATIONS AND APPLICATIONS

In this section, through extensive numerical examples, we show robustness and improved concentration about the population parameter of the geometric median of subset estimators in agreement with theorem 3.1. We first consider some simulated examples in estimating population means in  $S^d$  and  $PD(3)$ . We then formulate a robust procedure for estimating explanatory directions for dimension reduction in  $PD(3)$  and do a simulation study using this procedure. Finally, we apply the median-of-means method in the shape space to a hand shape data set as in [11].

Numerical results from both simulated and real data analysis in this section agree with the robustness and concentration properties of the estimator. We see in these results that

- (1) In simulations 1, 2, 3, and 4, and with various numbers of outliers, the average distance of the median-of-means is always an improvement over the average distances of the subset means.
- (2) The average distance of the median-of-means is almost always an improvement over the overall mean in the presence of outliers.
- (3) In the case of  $PD(3)$ , in Simulation 4, the average distance of the median-of-means for  $m = 5, 10, 15$  often gives an improvement over the overall median ( $m = 60$ ) in the presence of outliers. Number of groups  $m = 15$  seems to provide the best concentration overall. That the effect in more pronounced seems to agree with the log map in  $PD(3)$  being 1-Lipschitz as in proposition 2.3 and with the bound given in theorem 3.1 with  $K = 1$ .

In simulation 5, the median-of-means estimator is applied in estimating both the center of operations and explanatory directions for dimension reduction. The robustness property is shown as explanatory submanifolds maintain their fit to data in terms of intrinsic sum-of-squared residuals in the presence of outliers better than the ordinary PGA procedure. All code and data used in this section can be found in <https://github.com/DrewLazar/RobustManifold>.

**4.1. Simulation Study on  $S^d$ .** In this subsection, we provide examples with data simulated from the von Mises-Fisher distribution on the sphere. We consider the estimation of both intrinsic and extrinsic means in the presence of various numbers of outliers. As shown by the numerical comparisons below, the estimator obtained from the robust estimation procedure shows improved concentration over subset-based estimators and often is closer to the true parameter of interest compared to the overall sample mean and overall sample median. We first describe algorithms used for computing various summary statistics related to our estimators in  $S^d$ .

**4.1.1. Computation of Sample Statistics on  $S^d$ .** Given  $\{p_1, \dots, p_n\} \subset S^d$  we compute sample statistics as follows:

- (1) **Intrinsic mean.** With objective function  $L_n(x) = \frac{1}{n} \sum_{i=1}^n \arccos(\langle x, p_i \rangle)$  and constraint function  $g(x) = \langle x, x \rangle$  let

$$\gamma_i(x) = \frac{\arccos(\langle x, p_i \rangle)}{\sqrt{1 - \langle x, p_i \rangle^2}}.$$

Then the sample mean  $\hat{\mu}$  satisfies Lagrange multiplier condition

$$\sum_{i=1}^n \gamma_i(\hat{\mu}) p_i = \lambda \hat{\mu} \text{ with } \langle \hat{\mu}, \hat{\mu} \rangle = 1 \text{ and } \lambda = \sum_{i=1}^n \gamma_i(\hat{\mu}) \langle p_i, \hat{\mu} \rangle.$$

As in [13], letting  $\Psi(x) = \sum_{i=1}^n \gamma_i(x) p_i$ , we use the fixed-point algorithm

$$\begin{aligned} \mu_k &\mapsto \mu_{k+1} \\ \mu_{k+1} &= \frac{\Psi(\mu_k)}{\|\Psi(\mu_k)\|}. \end{aligned}$$

Then  $\mu_k \rightarrow \hat{\mu}$ .

- (2) **Intrinsic median.** We use a generalization of Ostresh's modification of Weiszfeld's algorithm as introduced in [11]. Let

$$\Psi(x) = \sum_i \frac{\text{Log}_x(p_i)}{\arccos(\langle x, p_i \rangle)} \left( \sum_i \frac{1}{\arccos(\langle x, p_i \rangle)} \right)^{-1}$$

$$\begin{aligned} m_k &\mapsto m_{k+1} \\ m_{k+1} &= \text{Exp}_{x_k}(\Psi(m_k)). \end{aligned}$$

Then  $m_k \rightarrow \hat{m}$  where  $\hat{m}$  is the intrinsic sample median.

- (3) **Extrinsic mean.** As in [5], the extrinsic sample mean is the projection of the sample mean under the embedding. That is,

$$\hat{\mu} = \mathcal{P} \left( \frac{1}{n} \sum_{i=1}^n J(p_i) \right)$$

where  $J$  is our embedding map. In the case of  $S^d$ , where  $J$  is the identity map and projection is done by normalizing in  $\mathbb{R}^{d+1}$ ,  $\mu = \bar{x} / \|\bar{x}\|$  where  $\bar{x}$  is the Euclidean sample mean.

- (4) **Extrinsic median.** Let

$$L_n(x) = \frac{1}{n} \sum_{i=1}^n \|x - p_i\| \text{ for } x \in \mathbb{R}^{d+1} \text{ and } g(x) = L_n|_{S^d}.$$

With  $S^d$  as a submanifold of  $\mathbb{R}^{d+1}$ , for  $p \in S^d$  the gradient of  $g$  is the orthogonal projection of  $\nabla_p L_n$  onto  $T_p S^d$ , that is,

$$\nabla_p g = \text{proj}_{T_p S^d}(\nabla_p L_n).$$

We take  $\nabla_p L_n$  as in Weiszfeld's algorithm [33] and compute the sample geometric median  $\hat{m}$  by gradient descent as follows:

$$\begin{aligned} \Psi(x) &= \sum_i \frac{p_i - x}{\|p_i - x\|} \left( \sum_i \frac{1}{\|p_i - x\|} \right)^{-1} \\ m_k &\mapsto m_{k+1} \\ m_{k+1} &= \text{Exp}_{m_k} \left( \text{proj}_{T_{m_k} S^d}(\Psi(m_k)) \right) \end{aligned}$$

Then  $m_k \rightarrow \hat{m}$ , the extrinsic sample median.

**4.1.2. Simulations in  $S^d$ .** We consider the von Mises-Fisher distribution on the unit sphere. Distributions on the sphere, and the estimations of their intrinsic means have important applications in directional statistics, as in [25], and cluster analysis, as in [3]. A von Mises-Fisher distribution on  $S^d$  has pdf

$$f_d(x; \mu, \kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} e^{\kappa \langle \mu, x \rangle},$$

where  $I_n$  is the modified Bessel function of the first kind

$$I_n(\kappa) = \frac{2^{-n} \kappa^n}{\Gamma(n+1/2)\Gamma(1/2)} \int_0^\pi e^{\kappa \cos \theta} \sin^{2n} \theta d\theta.$$

The intrinsic mean of the distribution is  $\mu$  and  $\kappa$  is a concentration parameter about  $\mu$  with a larger  $\kappa$  giving increased concentration. One has

$$f_d(x; \mu, \kappa) = \frac{\Gamma((d-1)/2)}{2\pi^{(d-1)/2} \int_0^\pi e^{\kappa \cos \theta} \sin^{d-2} \theta d\theta} e^{\kappa \langle \mu, x \rangle}.$$

Thus, sampling  $x$  from the von Mises-Fisher distribution,

$$\begin{aligned} P(d_g(x, \mu) \leq \epsilon) &= \frac{\Gamma((d-1)/2)}{2\pi^{(d-1)/2} \int_0^\pi e^{\kappa \cos \theta} \sin^{d-2} \theta d\theta} \left( \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \right) \int_0^\epsilon e^{\kappa \cos \theta} \sin^{d-2} \theta d\theta \\ (4.1) \quad &= \frac{\int_0^\epsilon e^{\kappa \cos \theta} \sin^{d-2} \theta d\theta}{\int_0^\pi e^{\kappa \cos \theta} \sin^{d-2} \theta d\theta}. \end{aligned}$$

*Simulation 1.* Estimating Intrinsic Mean in  $S^2$ : Using [14] we sample  $n = 60$  data points from the von Mises-Fisher distribution on  $S^2$ . We take  $\kappa = 30$ , which by (4.1) guarantees with probability  $\approx 1$  that the sample is within a hemisphere and thus the intrinsic mean and median uniquely exist.

We include  $k = 0, 5, 10$ , and 15 outliers outside a symmetric 95% confidence region about the mean with the confidence region computed using (4.1). We then apply the median-of-means technique of section 3 for  $m = 1, 5, 15, 30$  and 60 groups. Over 1000 runs, we compute

- (1) the average intrinsic distance  $\overline{\rho(\mu^*, \mu)}$  from the true mean  $\mu$  to the geometric median of subsets estimator  $\mu^*$ .

- (2) the average intrinsic distance  $\overline{\rho(\mu_i, \mu)}$  from  $\mu$  to the average of the subset means  $\mu_i$ ,  $i = 1, \dots, m$ .

| k  | $\overline{\rho(\hat{\mu}, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|----|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 0  | 0.0597                            | 0.0583                        | 0.0947                        | 0.0514                        | 0.1496                        |
| 5  | 0.0647                            | 0.0615                        | 0.1159                        | 0.0531                        | 0.1652                        |
| 10 | 0.1194                            | 0.1116                        | 0.1414                        | 0.1018                        | 0.2113                        |
| 15 | 0.1819                            | 0.1731                        | 0.1973                        | 0.1631                        | 0.2419                        |
|    | sample mean (m=1)                 | m=5                           |                               | m=15                          |                               |

| k  | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\hat{m}, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|----|-------------------------------|-------------------------------|---------------------------------|-------------------------------|
| 0  | 0.0455                        | 0.2118                        | 0.0424                          | 0.2829                        |
| 5  | 0.0453                        | 0.2350                        | 0.0447                          | 0.2959                        |
| 10 | 0.0776                        | 0.2501                        | 0.0614                          | 0.3259                        |
| 15 | 0.1383                        | 0.2954                        | 0.0925                          | 0.3738                        |
|    | m=30                          |                               | sample median (m=60)            |                               |

TABLE 1. Results from Simulation 1 showing performance for various estimators of the mean under a von Mises-Fisher distribution in  $S^2$ , with  $k$  the number of outliers and  $\rho$  intrinsic distance.

Note that when  $m = 1$ ,  $\mu_i$  and  $\mu^*$  are both the sample Fréchet mean of the whole data set, which we denote as  $\hat{\mu}$ . Also, when  $m = 60$ ,  $\mu^*$  is the sample median and  $\mu_i = p_i$  for  $i = 1, \dots, 60$ . The same situation holds in simulations 2, 3 and 4.

In Figure 2 we have a sample of  $n = 60$  from the von Mises-Fisher Distribution including 5 added outliers. We take  $m = 5$  subsets and we see the improved concentration about the population mean of the geometric median of the 5 subset means.

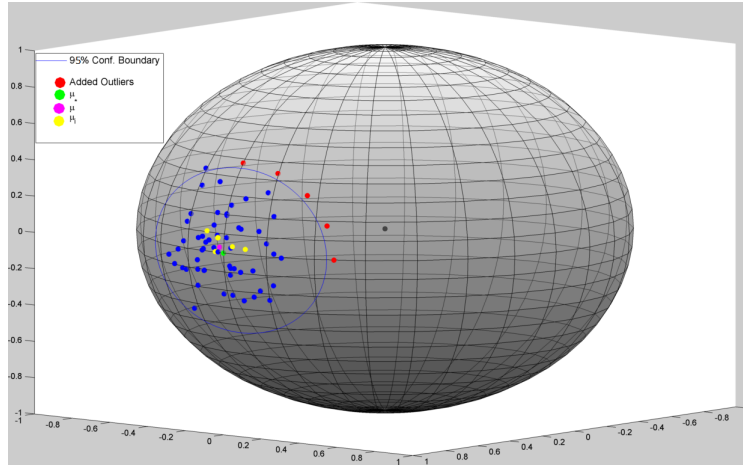


FIGURE 2. Von-Mises Fisher,  $\kappa = 30$ , 5 added outliers

*Simulation 2.* Approximation of the Intrinsic Mean in  $S^7$ : We repeat the first part of the experiment in Simulation 1 in  $S^7$  except with  $n = 200$ ,  $\kappa = 20$ ,  $k = 0, 10, 20, 40$  outliers and  $m = 1, 10, 50, 100, 200$  groups.

| k                 | $\overline{\rho(\hat{\mu}, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|-------------------|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 0                 | 0.0396                            | 0.0399                        | 0.1186                        | 0.0384                        | 0.2570                        |
| 10                | 0.0565                            | 0.0541                        | 0.1258                        | 0.0514                        | 0.2669                        |
| 20                | 0.0897                            | 0.0900                        | 0.1462                        | 0.0834                        | 0.2827                        |
| 40                | 0.1656                            | 0.1678                        | 0.2082                        | 0.1596                        | 0.3376                        |
| Sample mean (m=1) |                                   | m=10                          |                               | m=50                          |                               |

| k     | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\hat{m}, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|-------|-------------------------------|-------------------------------|---------------------------------|-------------------------------|
| 0     | 0.0398                        | 0.3590                        | 0.0387                          | 0.4896                        |
| 10    | 0.0469                        | 0.3676                        | 0.0457                          | 0.4978                        |
| 20    | 0.0760                        | 0.3896                        | 0.0682                          | 0.5301                        |
| 40    | 0.1513                        | 0.5176                        | 0.1305                          | 0.5987                        |
| m=100 |                               | sample median (m=200)         |                                 |                               |

TABLE 2. Results from Simulation 2 showing performance for various estimators of the mean under a von Mises-Fisher distribution in  $S^7$ , with  $k$  the number of outliers and  $\rho$  intrinsic distance.

*Simulation 3.* Approximation of the Extrinsic Mean in  $S^2$ : We repeat the experiment in Simulation 1, but with  $\rho$  as the extrinsic distance and with each average taken over 1200 runs.

| k                 | $\overline{\rho(\hat{\mu}, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|-------------------|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 0                 | 0.0272                            | 0.0330                        | 0.0676                        | 0.0312                        | 0.1179                        |
| 5                 | 0.0621                            | 0.0634                        | 0.0943                        | 0.0541                        | 0.1512                        |
| 10                | 0.1231                            | 0.1190                        | 0.1456                        | 0.1083                        | 0.1952                        |
| 15                | 0.1771                            | 0.1688                        | 0.1956                        | 0.1632                        | 0.2337                        |
| Sample mean (m=1) |                                   | m=5                           |                               | m=15                          |                               |

| k    | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\hat{m}, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|------|-------------------------------|-------------------------------|---------------------------------|-------------------------------|
| 0    | 0.0305                        | 0.1681                        | 0.0312                          | 0.2312                        |
| 5    | 0.0453                        | 0.2034                        | 0.0411                          | 0.2745                        |
| 10   | 0.0847                        | 0.2479                        | 0.0612                          | 0.3241                        |
| 15   | 0.1453                        | 0.2971                        | 0.0837                          | 0.3728                        |
| m=30 |                               | sample median (m=60)          |                                 |                               |

TABLE 3. Results from Simulation 3 showing performance for various estimators of the mean under a von Mises-Fisher distribution in  $S^7$ , with  $k$  the number of outliers and  $\rho$  intrinsic distance.

The results in Tables 1-3, showing the performance of the various estimators in Simulations 1-3 respectively, demonstrates that the median-of-means estimator almost always improves over the average of subset means and overall Fréchet sample mean estimators in the presence of outliers.

**4.2. Simulation study on  $PD(3)$ .** In this subsection, we consider simulated data from a generalized log-normal distribution on the space of  $3 \times 3$  positive definite matrices,  $PD(3)$ . As in subsection 4.1, we consider the estimation of intrinsic means in the presence of various numbers of outliers. There are multiple applications in which it is of interest to estimate the mean of a sample of positive definite matrices. This includes principal geodesic analysis (PGA), as in [10], where optimization to find explanatory directions is done in the tangent space at the sample mean. Using our median-of-means procedure, we formulate a robust PCA procedure (RPGA). We first describe algorithms used for computing various summary statistics related to our estimators in  $PD(3)$ .

**4.2.1. Computation of Sample Statistics on  $PD(3)$ .** To compute the sample intrinsic mean in the following simulation, we use the damped gradient descent algorithm as in [10]. As shown in [15], as  $PD(3)$  is of non-negative curvature, the intrinsic mean is guaranteed to exist and to be unique. To compute the sample intrinsic median, we use the generalization of Weiszfeld's algorithm given in [11] where the sample intrinsic median is shown to exist and to be unique. Computations of projection to subspaces and of principal geodesic directions are done using MATLAB minimization routines and user-supplied gradients as formulated in [32] with the derivative of the matrix exponential map provided by [28, Theorem 4.5].

**4.2.2. Robust Principal Geodesic Analysis (RPGA).** Principal Geodesic Analysis (PGA) as in [18] is a two-step procedure which involves 1) computing a center of the data and 2) successively finding orthogonal tangent vectors at that center so that their exponentiated span best fits the data according to intrinsic sum-of-squared residuals.

We propose a Robust PGA procedure (RPGA) which 1) uses the median-of-means estimate as the center of the data and 2) finds orthogonal directions in the tangent space using the robust median-of-means Principal Component Analysis (PCA) procedure given in [26]. Specifically, in RPGA

- (1) Divide the data into  $m$  subsets  $U_1, \dots, U_m$  and for each compute an intrinsic mean  $\mu_j$  as in (3.1) and then compute  $\mu^* = \text{med}(\mu_1, \dots, \mu_m)$  as in (3.2).
- (2) Compute  $V_i = \text{vec}(\text{Log}_{\mu^*}(U_i))$  where  $\text{Log}_{\mu^*}(U_i)$  is the image of  $U_i$  under the Riemmanian log map. As in [26], compute sample covariance matrices  $\Sigma_i$  for each  $V_i$  and then compute

$$\hat{\Sigma} = \text{med}(\Sigma_1, \dots, \Sigma_n)$$

where the median is taken with respect to Frobenius norm  $\|A\|_F = \text{trace}(A^T A)$ .

We take the eigenvectors of  $\hat{\Sigma}$ ,  $\{w_1, \dots, w_6\}$ , arranged in order by largest to smallest eigenvalue. Then our robust principal geodesic directions in the tangent space at  $\mu^*$  are  $\{v_1, \dots, v_6\}$  where  $v_i$  is the vector corresponding to  $w_i$  by the vec operator. To form explanatory subspaces we then exponentiate the span of  $\{v_1, \dots, v_k\}$  at  $\mu^*$  for  $k = 1, \dots, 6$ .

This procedure is robust as it ensures both the located center of the data and the located explanatory directions are not unduly affected by the presence of outliers.

#### 4.2.3. Simulations in $PD(3)$ .

*Simulation 4.* Estimating the Intrinsic Mean in  $PD(3)$ : We sample  $n = 60$  data points from a log-normal distribution where if the random variable  $X$  has this distribution then  $\text{vec}(\text{Log}_I(X)) \sim \mathcal{N}(\mathbf{0}, \kappa \mathbf{I})$  with  $\kappa$  a scaling parameter. We repeat the experiment of Simulation 1 of section 4.1.2 with each average taken over 1200 runs.

| k                 | $\overline{\rho(\hat{\mu}, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|-------------------|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 0                 | 0.2630                            | 0.2781                        | 0.5909                        | 0.2753                        | 1.0408                        |
| 5                 | 0.2640                            | 0.2512                        | 0.5776                        | 0.2683                        | 1.0745                        |
| 10                | 0.3568                            | 0.3179                        | 2.7485                        | 0.2986                        | 1.3158                        |
| 15                | 0.5292                            | 0.3001                        | 1.0433                        | 0.3437                        | 1.4246                        |
| Sample mean (m=1) |                                   | m=5                           |                               | m=15                          |                               |

| k    | $\overline{\rho(\mu^*, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ | $\overline{\rho(\hat{m}, \mu)}$ | $\overline{\rho(\mu_i, \mu)}$ |
|------|-------------------------------|-------------------------------|---------------------------------|-------------------------------|
| 0    | 0.2750                        | 1.5230                        | 0.2728                          | 2.3449                        |
| 5    | 0.2724                        | 1.5930                        | 0.2675                          | 2.4139                        |
| 10   | 0.3306                        | 1.7607                        | 0.3482                          | 2.5002                        |
| 15   | 0.4183                        | 1.8107                        | 0.5265                          | 2.5617                        |
| m=30 |                               | Sample median (m=60)          |                                 |                               |

TABLE 4. Results for Simulation 4 with data simulated from a log-normal distribution in  $PD(3)$ ,  $k$  the number of outliers, and  $\rho$  the intrinsic distance.

The results are shown in Table 4. Again, in this example, the median-of-mean estimator always improves over the average of the means and almost always over the overall sample Fréchet mean. The average distance from the truth of the median-of-means for  $m = 5, 10, 15$  is an improvement over the overall median ( $m = 60$ ) in the presence of outliers. The number of groups  $m = 15$  seems to provide the best concentration overall.

*Simulation 5.* Estimating Explanatory Directions in  $PD(3)$  with RPGA: We sample from a log-normal distribution, where if the random variable  $X$  has this distribution then  $\text{vec}(\text{Log}_I(X)) \sim \mathcal{N}(\mathbf{0}, \kappa \Sigma)$  with  $\kappa$  a scaling parameter.  $\Sigma$  is diagonal with diagonal entries which vary from 1 to 20 to ensure that population PGA directions exist.

Over 200 runs, we add 0, 5, 10, 15 outliers outside a 95% confidence region in  $n = 60$  data points and compute PGA and RPGA explanatory directions. We then find the intrinsic mean sum of squared residuals (mSSRs) of the data without outliers relative to the estimated explanatory submanifolds. Table 5 gives the average of the mSSRs over 200 runs for submanifolds of 1, 2, and 3 dimensions for PGA and for RPGA computed with 5, 10 and 15 groups.

We see that without outliers, the PGA procedure, which sequentially optimizes a fit to the data at the intrinsic mean, produces the lowest average mSSR, regardless of the number of groups for RPGA. However, as outliers are added, the mSSR

for PGA increases to a greater extent than RPGA. Note that RPGA with  $m = 1$  groups is the linear approximation of the PGA procedure given in [10]

| k          | PGA    | RPGA   | RPGA   | RPGA   |
|------------|--------|--------|--------|--------|
| 0          | 0.4206 | 0.4265 | 0.4259 | 0.4320 |
| 5          | 0.4529 | 0.4465 | 0.4314 | 0.4342 |
| 10         | 0.4541 | 0.4438 | 0.4508 | 0.4374 |
| 15         | 0.4540 | 0.4445 | 0.4492 | 0.4442 |
| 20         | 0.4527 | 0.4473 | 0.4507 | 0.4496 |
| $m$ groups |        | m=5    | m=10   | m=15   |

| k          | PGA    | RPGA   | RPGA   | RPGA   |
|------------|--------|--------|--------|--------|
| 0          | 0.2629 | 0.2686 | 0.2691 | 0.2751 |
| 5          | 0.2924 | 0.2870 | 0.2803 | 0.2795 |
| 10         | 0.2963 | 0.2838 | 0.2925 | 0.2791 |
| 15         | 0.2994 | 0.2835 | 0.2758 | 0.2850 |
| 20         | 0.3041 | 0.2841 | 0.2889 | 0.2775 |
| $m$ groups |        | m=5    | m=10   | m=15   |

| k          | PGA    | RPGA   | RPGA   | RPGA   |
|------------|--------|--------|--------|--------|
| 0          | 0.1472 | 0.1497 | 0.1533 | 0.1608 |
| 5          | 0.1919 | 0.1801 | 0.1600 | 0.1588 |
| 10         | 0.2242 | 0.2102 | 0.1940 | 0.1743 |
| 15         | 0.2208 | 0.2149 | 0.2134 | 0.2079 |
| 20         | 0.2305 | 0.2259 | 0.2169 | 0.2206 |
| $m$ groups |        | m=5    | m=10   | m=15   |

TABLE 5. Average mSSRs to explanatory submanifolds computed with  $k$  outliers to data without outliers in  $PD(3)$

**4.3. Hand Shape Data in  $\Sigma_2^K$ .** We consider the hand shape data set in [8] of 18 hands with each hand in planar shape space  $\Sigma_2^{72}$ . A planar shape  $\Sigma_2^K$  consists of objects with  $K$  landmarks in  $\mathbb{R}^2$  modulo the Euclidean motions including rotation, scaling and translation [5, 16]. As in [11], we use ellipses as outliers with each one as

$$\{(a \cos(k\pi/36), b \sin(k\pi/36)); k = 0, \dots, 71\}$$

where  $a, b$  are sampled from the uniform distribution on  $[0.5, 1]$ . With  $k = 3$  added outliers, we divide the data of size  $n = 21$  into  $m = 7$  random subsets, each of size 3. We then compute and observe the geometric median and the sample mean.

**4.3.1. Computation of Sample Statistics on  $\Sigma_2^K$ .** We identify  $\Sigma_2^{72}$  with  $S^{69}/S^1$  as in (2.9), and compute intrinsic sample means and medians using direct modifications of the algorithms in section 4.1.1.

In Figure 3 (a) we show  $n = 21$  hands with 3 outliers. In (b) we show 7 randomly assigned subsets indicated with seven different colors, and in (c) we show the subset means of each group. In (d) we see less influence of the outliers in the geometric median, as it retains the shape of a hand similar to the original 18 hands.

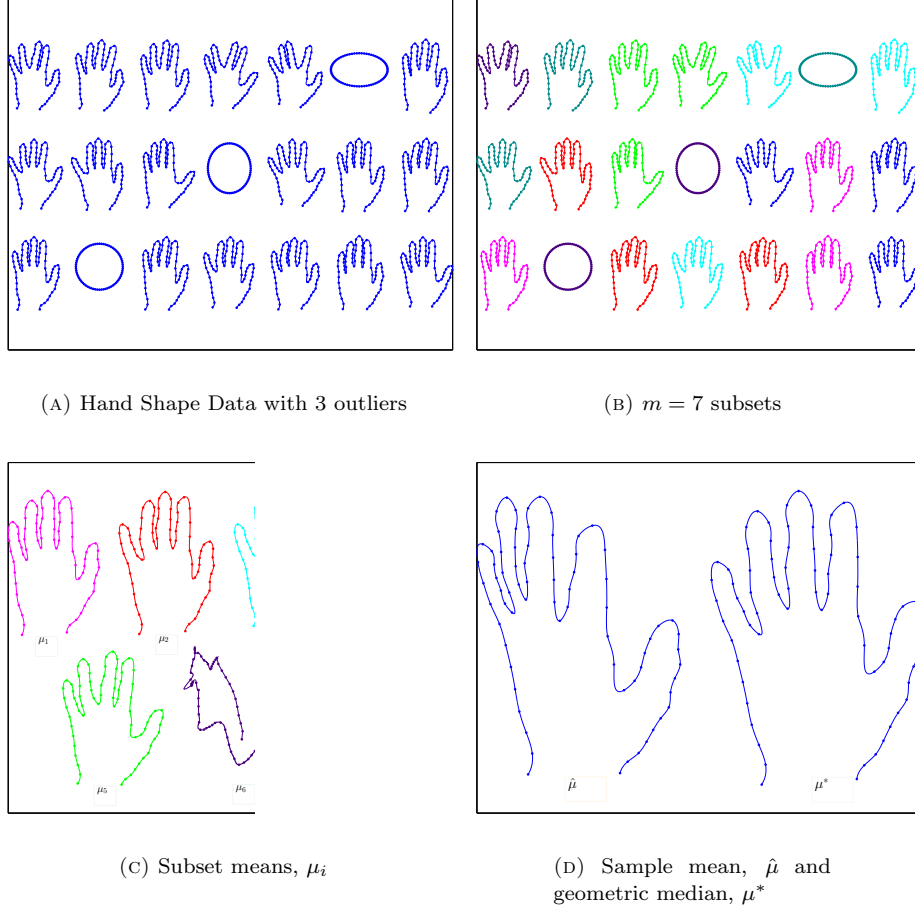


FIGURE 3. Median-of-Means on Hand Shape Data

## 5. DISCUSSION

We propose a robust and scalable procedure for general optimization problems on manifolds. Scalability is of particular importance in dealing with the difficult computational issues that arise in estimating sample statistics for manifold data or extracting low-dimensional manifold in high-dimensional data. Along these lines, parallel computation can be implemented trivially from the subsampling procedure.

It is shown through lemma 2.1, which provides an important property of geometric medians on manifolds, and the following theorem 3.1, that the resulting estimator yields provable robustness and tighter concentration bounds about the true parameter of interest. Numerical results from both simulated and real data analysis in Section 4 agree with the robustness and concentration properties of the estimator.

Future research might include considering the optimal numbers and sizes of subgroups for estimation as discussed in remark 3.1. In theorem 3.1, for a given  $\epsilon$ , more groups provide a larger  $m$  but also a larger  $\eta$  in the bound provided by 3.4. This

is reflected in Examples 1 and 2 in (3.6) and (3.7). In finding the optimal  $m$ , the number of outliers and amount of contamination in the data also must be factored in. In addition, challenging computational considerations for large data sets on manifolds and advantages in partitioning data needs to be considered. Also, the second step of the RPGA procedure in 4.2.2 might be done instead by partitioning the data in the manifold rather than their Riemannian logs in the tangent space at  $\mu^*$ . Computation of RPGA, as formulated in 4.2.2, only requires the computation of the median-of-means  $\mu^*$ , and then the linear operation of computing sample covariance matrices of the Riemannian logs of data in the tangent space at  $\mu^*$ . Robust estimation on manifolds in other contexts such as manifold regression [2] might also be considered. As in the case of estimation of the mean, additional machinery and complications arise in the more general context of a manifold.

#### ACKNOWLEDGEMENTS

Lizhen Lin would like to thank Dong Quan Nguyen for very useful discussions. The contribution of LL and BS was funded by NSF grants IIS 1663870, DMS CAREER 1654579 and a DARPA grant N66001-17-1-4041.

#### REFERENCES

- [1] Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S. Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.
- [2] Anil Aswani, Peter Bickel, Claire Tomlin, et al. Regression on manifolds: estimation of the exterior derivative. *The Annals of Statistics*, 39(1):48–81, 2011.
- [3] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.
- [4] Rajendra Bhatia. On the exponential metric increasing property. *Linear Algebra and its applications*, 375:211–220, 2003.
- [5] A. Bhattacharya and R. Bhattacharya. *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*. IMS Monograph #2. Cambridge University Press, 2012.
- [6] R. Bhattacharya and L. Lin. An omnibus CLT for Fréchet means and nonparametric inference on non-Euclidean spaces. *ArXiv eprint*, 1306.5806, June 2013.
- [7] Nicolas Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications*, 475:200 – 239, 2015.
- [8] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995.
- [9] W. Dai, E. Kerman, and O. Milenkovic. A geometric approach to low-rank matrix completion. *IEEE Transactions on Information Theory*, 58(1):237–247, 2012.
- [10] P. Thomas Fletcher and Sarang Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.*, 87(2):250–262, February 2007.
- [11] P. Thomas Fletcher, Suresh Venkatasubramanian, and Sarang C. Joshi. Robust statistics on riemannian manifolds via the geometric median. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 24–26 June 2008, Anchorage, Alaska, USA, 2008.
- [12] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948.
- [13] Stephan Huckemann and Herbert Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.*, 38(2):299–319, 2006.
- [14] S. Jung. Random number generatrion form von mises-fisher distribution. *Technical report*, University of Pittsburgh, 2010.
- [15] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30(5):509–541, 1977.

- [16] D. G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. of the London Math. Soc.*, 16:81–121, 1984.
- [17] Eric D. Kolaczyk, Lizhen Lin, Steven Rosenberg, Jackson Walters, and Jie Xu. Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *Ann. Statist.*, 48(1):514–538, 02 2020.
- [18] Drew Lazar and Lizhen Lin. Scale and curvature effects in principal geodesic analysis. *Journal of Multivariate Analysis*, 153:64–82, 2017.
- [19] Guillaume Lecu and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.*, 48(2):906–931, 04 2020.
- [20] M. Lerasle and R. I. Oliveira. Robust empirical mean Estimators. *ArXiv e-prints*, December 2011.
- [21] Lizhen Lin, Vinayak Rao, and David B. Dunson. Bayesian nonparametric inference on the Stiefel manifold. *Statistics Sinica*, 27:535–553, 2017.
- [22] Lizhen Lin, Brian St. Thomas, Hongtu Zhu, and David B. Dunson. Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association*, 112(519):1261–1273, 2017.
- [23] Suhas Lohit and Pavan K. Turaga. Learning invariant riemannian geometric representations using deep nets. *2017 IEEE International Conference on Computer Vision Workshops (IC-CVW)*, pages 1329–1338, 2017.
- [24] Gbor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 08 2019.
- [25] K. V. Mardia. *Statistics of directional data*. Academic Press, London-New York, 1972. Probability and Mathematical Statistics, No. 13.
- [26] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 11 2015.
- [27] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18(124):1–40, 2017.
- [28] Igor Najfeld and Timothy F Havel. Derivatives of the matrix exponential and their computation. *Advances in applied mathematics*, 16(3):321–375, 1995.
- [29] A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization, 1983. unpublished.
- [30] S. P. Novikov and I. A. Taimanov. *Modern Geometric Structures and Fields. Transl. from the Russian by D. Chibisov.*, volume 71. Providence, RI: American Mathematical Society (AMS), 2006.
- [31] Bayan Saporbayeva, Michael Minyi Zhang, and Lizhen Lin. Communication efficient parallel algorithms for optimization on manifolds. In *NeurIPS*, 2018.
- [32] Stefan Sommer, François Lauze, and Mads Nielsen. The differential of the exponential map, jacobi fields and exact principal geodesic analysis. *CoRR*, abs/1008.1902, 2010.
- [33] E. Weiszfeld. On the point for which the sum of the distances to  $n$  given points is minimum. *Ann. Oper. Res.*, 167:7–41, 2009. Translated from the French original [Tohoku Math. J. 43 (1937), 355–386] and annotated by Frank Plastria.

DEPARTMENT OF APPLIED AND COMPUTATIONAL MATHEMATICS AND STATISTICS, THE UNIVERSITY OF NOTRE DAME, NOTRE DAME, IN.

*E-mail address:* lizhen.lin@nd.edu