Statistica Sinica **34** (2024), 1299-1323
doi:https://doi.org/10.5705/ss.202022.0149

ROBUST OPTIMIZATION AND INFERENCE ON MANIFOLDS

Lizhen Lin*, Drew Lazar, Bayan Saparbayeva and David Dunson

The University of Maryland, Ball State University, University of Rochester and Duke University

Abstract: We propose a robust and scalable procedure for general optimization and inference problems on manifolds, leveraging the classic 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 the 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 using examples. The robustness and scalability of the procedure is shown in numerical examples on simulated and real data sets.

Key words and phrases: 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 nonlinear Riemannian manifolds. Advancements in technology and computing have led to an increasing prevalence of complex data in non-Euclidean forms, such as positive-definite matrices (diffusion matrices) in diffusion tensor imaging (Alexander et al. (2007)), shape objects in medical vision (Kendall (1984)), network data objects (Kolaczyk et al. (2020)) and subspaces or orthonormal frames (Lin, Rao and Dunson (2017)). A proper statistical inference from such data involves optimizing over the underlying manifold to which the data are constrained. For example, there is a vibrant line of research on estimating Fréchet means (Fréchet (1948)), which are minimizers of Fréchet functions on manifolds (Bhattacharya and Bhattacharya (2012); Bhattacharya and Lin (2017)). In this case, both the data and the parameters of interest

*Corresponding author.

are on manifolds. In addition, it is common to represent a lower-dimensional structure in high-dimensional data as a manifold. Learning such a manifold is a nontrivial optimization problem. In each of the above problems, we require algorithms that are robust to data contamination and heavy tails and that scale efficiently to large data sets.

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 (Nemirovskij and Yudin (1983)) to manifolds by establishing some key properties of the geometric median on manifolds, with which we can prove the tighter concentration bounds of our proposed estimator. The key idea is to obtain optimizers from a subset of the data, aggregating them to form a final estimator. Our estimator is robust to outliers and contaminations of an arbitrary nature, and has provable robustness. The scalability of the algorithm is guaranteed by the divide-and-conquer nature of combining subset-based estimators.

There is a related body of literature outside the non-Euclidean manifold setting. For example, Minsker (2015) applies the median-of-means procedure for a robust estimation in Banach spaces. Minsker et al. (2017) and Minsker et al. (2014) propose a robust Bayesian estimator as the geometric median of measures of the subset posteriors. Characterizing the properties of the geometric median on manifolds requires a substantially different approach to deal with the underlying geometry. We prove a key lemma characterizing the robustness property of geometric medians on manifolds, which allows us to show that our estimator has tighter concentration bounds than those of 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 latter adopting a Riemannian structure. We illustrate the bounds with explicit calculations in both the extrinsic and the intrinsic cases. Our procedure is demonstrated in a class of manifolds using simulated and real-data examples. The manifolds we consider include the sphere, positive-definite matrices, and planar shape spaces, all of which are commonly applicable in real-data analysis.

The remainder of 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 when estimating the population parameter of interest, and provide examples of calculations of the bounds. In Section 4, a simulation study and data analysis are used to show the robustness and scalability of our procedure. The final section concludes the paper.

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*

$$\mu = \operatorname*{argmin}_{p \in \mathcal{M}} L^*(p), \tag{2.1}$$

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

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

for some loss function L. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, where $x_1, \ldots, x_n$ are sampled from Q. The parameter μ is often estimated using the *empirical risk estimator*

$$\hat{\mu}_n = \operatorname*{argmin}_{p \in \mathcal{M}} L_n(p, \boldsymbol{x}) = \operatorname*{argmin}_{p \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} L(p, x_i). \tag{2.2}$$

Remark 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 ρ a metric defined on $\mathcal{M}$. There is a significant amount of literature on nonparametric statistical inference on manifolds, in which the estimation of the Fréchet mean is addressed (see Bhattacharya and Bhattacharya (2012); Bhattacharya and Lin (2017)). Similarly, in a regression problem with manifold-valued output, the underlying problem can be cast as an optimization problem on manifolds (Lin et al. (2017)). In many other applications, we do not have $\mathcal{X} = \mathcal{M}$, with $\mathcal{X}$ a higher-dimensional ambient space, and the optimization is done over a lower-dimensional manifold, such as the Grassmannian (Lohit and Turaga (2017); Saparbayeva, Zhang and Lin (2018)), which has abundant applications in manifold learning and low-rank estimation matrix problems (Dai, Kerman and Milenkovic (2012); Boumal and Absil (2015)).

Real data sets often contain outliers, which may be errors, extreme observations, or contamination when sampling from heavy-tailed or mixture distributions. Thus, there is interest in a robust estimation of the population parameters, using estimators that are stable and not unduly affected 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 $p_1, \ldots, p_m \in \mathcal{M}$ minimizes the sum of the distances to the points, that is,

$$p^* = \text{med}(p_1, \ldots, p_m) = \underset{p \in \mathcal{M}}{\text{argmin}} \frac{1}{m} \sum_{k=1}^{m} \rho(p, p_k), \tag{2.3}$$

assuming that p^* exists and is unique. When $\mathcal{M}$ is a manifold, there are different ways to metrize the space. Let $J : \mathcal{M} \to \mathbb{R}^D$ be an embedding of a manifold $\mathcal{M}$ into some higher-dimensional Euclidean space $\mathbb{R}^D$. We denote the image of $\mathcal{M}$ after the embedding J as $\tilde{\mathcal{M}}$. That is, $\tilde{\mathcal{M}} = J(M)$. Note that $\tilde{\mathcal{M}}$ is a submanifold in $\mathbb{R}^D$. One can define an *extrinsic distance* on M induced from the embedding J using the Euclidean distance on $\tilde{\mathcal{M}}$. That is,

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

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^D$.

Alternatively, one can take ρ to be the *intrinsic distance*, as the geodesic distance arising from a Riemannian structure on $\mathcal{M}$. With the choice of ρ 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 Fletcher, Venkatasubramanian and Joshi (2008) by, for example, characterizing the uniqueness conditions of the intrinsic sample median, along with a Weiszfeld algorithm for finding the median. Our theoretical results 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 states that if $\omega \in \mathcal{M}$ is at least a constant, C_α times ϵ, distance away from the geometric median, p^*, then ω is at least ϵ distance away from at least an α fraction of the points $p_1, \ldots, p_m$. This result is illustrated in Figure 1. A similar result is proved in Minsker (2015) for Banach spaces. The proof of the following, a general lemma for manifolds, requires additional notation.

Lemma 1. *Let $p_1, \ldots, p_m \in \mathcal{M}$ and $p^* = \text{med}(p_1, \ldots, p_m)$, as in (2.3). Then, (a) and (b) hold:*

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

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

where $\alpha \in (0, \cot\psi \tan(\psi/2))$. If $\rho(\omega, p^*) \geq C_\alpha \epsilon$, then there exists an α portion of elements of $p_1, \ldots, p_m$ that are at least ϵ distance away from ω. That is, there exists an index set $T \subset \{1, \ldots, m\}$, with $|T| \geq \alpha m$, and

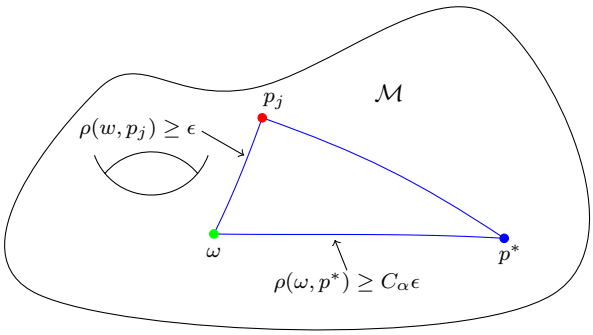


Figure 1. Geometric Illustration of Lemma 1 on Manifold $\mathcal{M}$.

$\rho(p_j, \omega) \geq \epsilon$, *for any* $j \in T$.

(b) Let ρ *be an intrinsic distance on* M *with respect to some Riemannian structure. Let* $\omega \in \mathcal{M}$, *and let the* log *map,* $\log_{p^*}$, *that is, the inverse exponential map* $\log_{p^*} = \exp_{p^*}^{-1}$, *be* K*-Lipschitz continuous from* $B(\omega, \epsilon)$ *to* $T_{p^*}\mathcal{M}$, *where the distance on* $T_{p^*}\mathcal{M}$ *is the Euclidean distance, 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* α *portion of elements of* $p_1, \ldots, p_m$ *that are at least* ϵ *distance away from* ω.

A detailed proof of Lemma 1 can be found in the Appendix. The key ideas to consider are the directional derivative of the objective function at the geometric median, and the "angle" between the curve connecting p^* and w and those connecting p^* and p_j. Assuming that the lemma does not hold leads to the directional derivative at the median being negative, contradicting the shown fact that we know that it is positive. At the same time, the directional derivatives, curves, and angles depend on whether we are using the extrinsic or intrinsic distance. For example, for the extrinsic case, the directional derivative of the objective function is defined along the curve that connects $J(p^*)$ and $J(w)$ on the image of the manifold $\tilde{\mathcal{M}}$.

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

Proposition 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, *given by*

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

is 2*-Lipschitz continuous from* $B(p,\pi/2)$ *to* T_pS^d *for all* $p \in S^d$.

The following proposition shows that the log map in similarity shape spaces Kendall (1984) also satisfies the K-Lipschitz condition.

Proposition 2. *The similarity or planar shape space is given as*

$$\Sigma_2^k = \frac{S^{2k-3}}{S^1}. \tag{2.4}$$

The inverse exponential map, $\log_p$*, given by*

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

on Σ_2^k *is* 2*-Lipschitz continuous from* $B(p,\pi/4)$ *to* $T_p\Sigma_2^k$*, for all* $p \in \Sigma_2^k$.

Proposition 3. *The manifold of positive-definite* n*-by-*n *matrices,* $PD(n)$*, has a* 1*-Lipchitz continuous inverse exponential map at any* $p \in \mathrm{PD}(n)$*. For a given metric, we have the following exponential and logarithm mappings:*

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

where

$$\exp X = I + \frac{Y}{1!} + \frac{Y^2}{2!} + \cdots + \frac{Y^n}{n!} + \cdots,$$
$$\log x = (x-I) - \frac{(x-I)^2}{2} + \cdots + (-1)^{n-1}\frac{(x-I)^n}{n} + \cdots,$$

for any $A, X \in \mathrm{Sym}(n)$ *and any* $p, q, x \in \mathrm{PD}(n)$.

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 $x_1, \ldots, x_n$ into m subsets $U_1, \ldots, U_m$, each of roughly size $\lfloor n/m \rfloor$. Let $\mu_1, \ldots, \mu_m$ be the optimizers of the empirical risk function from each subset, $U_1, \ldots, U_m$, respectively. That is,

$$\mu_j = \operatorname*{argmin}_{p \in \mathcal{M}} L_{|U_j|}(p, U_j) \text{ for } j = 1, \ldots, m, \tag{3.1}$$

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

$$\mu^* = \operatorname*{argmin}_{p \in M} \sum_{j=1}^{m} \rho(p, \mu_j). \tag{3.2}$$

We show that μ^* exhibits robustness properties when estimating the population parameter μ.

Minsker (2015) proves that the geometric median of a collection of weakly concentrated estimators admits a tighter deviation bound in a Hilbert space. With the help of Lemma 1, we generalise this result to manifolds in the following theorem.

Theorem 1. *Let $\mu_1, \ldots, \mu_m$ be a collection of independent estimators of the parameter μ, and let the geometric median $\mu^* = \operatorname{med}(\mu_1, \ldots, \mu_m)$.*

(a) Let ρ be the extrinsic distance on $\mathcal{M}$ for some embedding $J : \mathcal{M} \to \tilde{\mathcal{M}} \subset \mathbb{R}^D$. Assume that 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 $\alpha \in (0, \cot\bar{\psi}\tan(\bar{\psi}/2))$, set

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

(b) Let ρ 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 $\alpha \in (0, 1/2)$, set

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

Under (a) or (b), if

$$P(\rho(\mu_j, \mu) > \epsilon) \le \eta \text{ for } i = 1, \ldots, n, \tag{3.3}$$

where $\eta < \alpha$, then

$$P(\rho(\mu^*, \mu) > \overline{C}_\alpha \epsilon) \le \exp(-m\phi(\alpha, \eta)), \tag{3.4}$$

where

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

Remark 2. An important aspect of constructing the estimator μ^* is the choice of the number of subsets m. By (3.4), a larger number of subset estimators yields greater robustness and a tighter concentration around the true parameter. However, there must be enough data in each subset to ensure that each subset estimator behaves well and η in (3.3) is sufficiently small. For a given confidence

level, ϵ, one can determine the number of subsets to achieve η 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 the extrinsic cases, of determining η in (3.3) that allows the computation of the bound in (3.4).

Example 1. Consider the embedding $J : \mathcal{M} \to \mathbb{R}^D$. We have the induced measure $\tilde{Q}$ on the image, where $\tilde{Q} = Q \circ J^{-1}$. Let $x_1, \ldots, x_n$ be an independent and identically distributed (i.i.d.) sample from a distribution Q, such that we have the extrinsic mean μ 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 $x_1, \ldots, x_n$ into m disjoint groups $U_1, \ldots, U_m$, each of size $[n/m]$, and define

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

Thus, we have 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,

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

By Chebyshev's inequality,

$$P\big(\rho(\mu_j, \mu) \geq \epsilon\big) = P\big(\rho^2(\mu_j, \mu) \geq \epsilon^2\big) \leq \frac{1}{\epsilon^2}\mathbb{E}\rho^2(\mu_j, \mu) \leq \frac{4}{\epsilon^2}\left[\frac{m}{n}\right]\mathbb{E}\rho^2(\mu, x_1). \quad (3.5)$$

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

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

where $\eta = (4/\epsilon^2)\,[m/n]\,\mathbb{E}\rho^2(\mu, x_1)$. Thus, by Theorem 1, for any $\alpha \in (0, \cot\bar{\psi}\ \tan(\bar{\psi}/2))$,

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

where

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

Example 2. Let $x_1, \ldots, x_n$ be an i.i.d. sample from a distribution Q, such that we have the Fréchet mean μ for the random variable x_1. Divide the sample $x_1, \ldots, x_n$ into m disjoint groups $U_1, \ldots, U_m$, each of size $[n/m]$, and define

$$\mu_j = \operatorname*{argmin}_{y \in \mathcal{M}} \frac{1}{|U_j|} \sum_{i \in U_j} d_g^2(y, x_i), \quad j = 1, \ldots, m.$$

Considering the jth subsample corresponding to U_j on the tangent space at μ_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 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,

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

Therefore, by Chebyshev's inequality,

$$P\big(d_g(\mu_j, \mu) \geq \epsilon\big) = P\big(d_g^2(\mu_j, \mu) \geq \epsilon^2\big) \leq \frac{1}{\epsilon^2}\mathbb{E}d_g^2(\mu_j, \mu) \leq \frac{K^2}{\epsilon^2}\left[\frac{m}{n}\right]\mathbb{E}d_g^2(\mu, x_1). \quad (3.6)$$

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

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

where $\eta = K^2 \, [m/n] \, \mathbb{E} d_g^2(\mu, x_1)$. Thus, by Theorem 1, for any $\alpha \in (0, 1/2)$,

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

where

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

Remark 3. Our proposed median-of-means estimator, is both robust and *scalable* over large data sets. For example, dividing the data into $m = 2, ..., \lfloor n/2 \rfloor$ subsets avoids expensive gradient descent steps computed over the entire data set when finding an overall sample median ($m = n$) or an overall sample mean ($m = 1$). Parallel processing can also be applied to compute the subset estimates simultaneously.

4. Simulations and Applications

In this section, using extensive numerical examples, we show the robustness and the improved concentration about the population parameter of the geometric median of subset estimators, supporting Theorem 1. We first consider simulated examples to estimate the 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 conduct 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 Fletcher, Venkatasubramanian and Joshi (2008).

The numerical results from the analyses of the simulated and real data presented in this section, agree with the robustness and concentration properties of the estimator. The results indicate the following:

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 improves on the overall median ($m = 60$)

in the presence of outliers. The number of groups $m = 15$ seems to provide the best concentration overall. That the effect is more pronounced seems to agree with the log map in $PD(3)$ being 1-Lipschitz, as in Proposition 3 and with the bound given in Theorem 1 with $K = 1$.

In simulation 5 we apply the median-of-means estimator to estimate the center of operations and explanatory directions for dimension reduction. The robustness property follows, because explanatory submanifolds maintain their fit to data in terms of the intrinsic sum-of-squared residuals in the presence of outliers better than the ordinary PGA procedure does. All code and data used in this section are available at `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 estimate 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 that of the subset-based estimators, and is often closer to the true parameter of interest than are the overall sample mean and the overall sample median. The algorithms used to compute the summary statistics related to our estimators in S^d, including the intrinsic mean, extrinsic mean, intrinsic median, and extrinsic median, are provided in the Supplementary Material.

4.1.1. Simulations in S^d

Simulation 1. Estimating the Intrinsic Mean in S^2: Following Jung (2010), we sample $n = 60$ data points from the von Mises–Fisher distribution on S^2. We take the concentration parameter $\kappa = 30$, which guarantees with probability ≈ 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. We then apply proposed median-of-means technique for $m = 1, 5, 15, 30$ and 60 groups. Over 1,000 runs, we compute the following:

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

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

Note that when $m = 1$, μ_i and μ^* are both the sample Fréchet mean of the whole data set, which we denote as $\hat{\mu}$. Furthermore, when $m = 60, \mu^*$ is

Table 1. Results from Simulation 1 showing the performance for various estimators of the mean under a von Mises–Fisher distribution in S^2, with k the number of outliers, and ρ the intrinsic distance.

k	$\overline{\rho(\hat{\mu},\mu)}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\hat{m},\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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)	

the sample median and $\mu_i = p_i$ for $i = 1, \ldots, 60$. 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 five added outliers. We take $m = 5$ subsets. The results show improved concentration about the population mean of the geometric median of the five subset means.

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.

Simulation 3. Approximation of the Extrinsic Mean in S^2: We repeat the experiment in Simulation 1, but with ρ as the extrinsic distance and with each average taken over 1,200 runs.

The results in Tables 1–3, showing the performance of the various estimators in Simulations 1–3 respectively, demonstrate that the median-of-means estimator almost always improves on the average of the subset means and on the 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 3×3 positive-definite matrices, $PD(3)$. As in subsection 4.1, we estimate intrinsic means in the presence of various numbers of outliers. Numerous applications seek to estimate the mean of a sample of positive-definite matrices, including a principal geodesic analysis (PGA), as in Fletcher and Joshi (2007), where the optimization to find explanatory directions occurs in

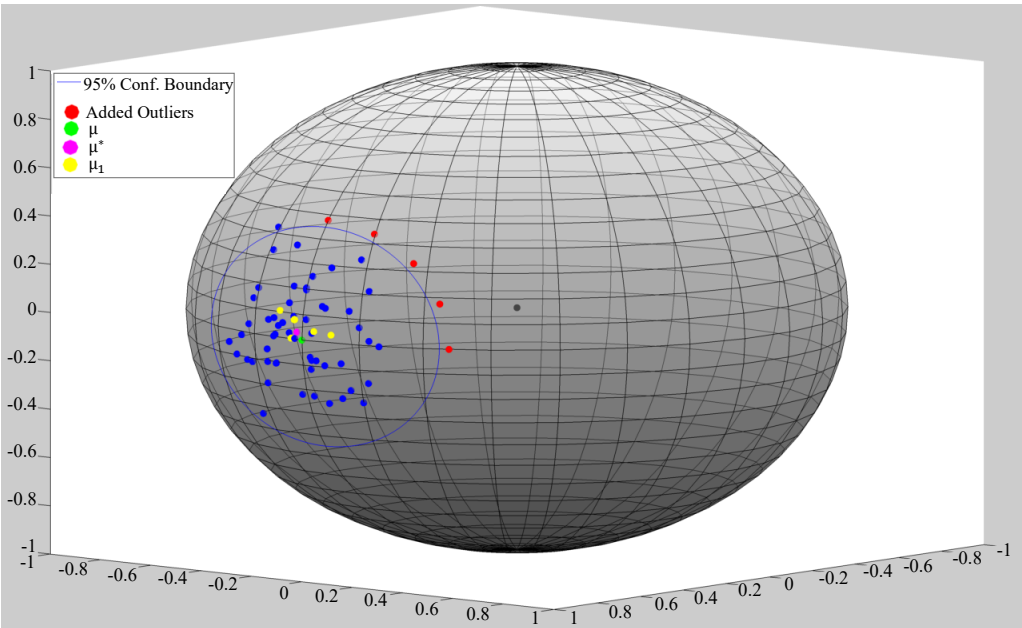


Figure 2. Von Mises–Fisher, $\kappa = 30$, five added outliers.

Table 2. Results from Simulation 2 showing the performance for various estimators of the mean under a von Mises–Fisher distribution in S^7, with k the number of outliers and ρ the intrinsic distance.

k	$\overline{\rho(\hat{\mu}, \mu)}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
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	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\hat{m}, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
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)	

the tangent space at the sample mean. Using our median-of-means procedure, we formulate a robust PCA procedure (RPGA). We first describe the algorithms used to compute the various summary statistics related to our estimators in $PD(3)$.

Table 3. Results from Simulation 3 showing the performance for various estimators of the mean under a von Mises–Fisher distribution in S^7, with k the number of outliers and ρ the intrinsic distance.

k	$\overline{\rho(\hat{\mu},\mu)}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\hat{m},\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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)	

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 Fletcher and Joshi (2007). As shown in Karcher (1977), because $PD(3)$ exhibits nonpositive curvature, the intrinsic mean is guaranteed to exist and be unique. To compute the sample intrinsic median, we use the generalization of Weiszfeld's algorithm given in Fletcher, Venkatasubramanian and Joshi (2008), where the sample intrinsic median is shown to exist and to be unique. The computations for the projection onto subspaces and the principal geodesic directions are performed using MATLAB minimization routines and user-supplied gradients, as in Sommer, Lauze and Nielsen (2010), with the derivative of the matrix exponential map provided by Najfeld and Havel (1995, Thm. 4.5).

4.2.2. Robust principal geodesic analysis

A principal geodesic analysis (PGA), as in Lazar and Lin (2017), is a two-step procedure that 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 the intrinsic sum-of-squared residuals.

We propose a robust PGA (RPGA) procedure that 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 of Minsker (2015). Specifically, in the RPGA, we do the following:

1. Divide the data into m subsets $U_1, \ldots, U_m$, and for each subset, compute an intrinsic mean μ_j, as in (3.1), and then compute $\mu^* = \text{med}(\mu_1, \ldots, \mu_m)$, as in (3.2).

2. Compute $V_i = \text{vec}(\text{L}og_{\mu^*}(U_i))$, where $\text{L}og_{\mu^*}(U_i)$ is the image of U_i under the Riemmanian log map. As in Minsker (2015), compute sample covariance matrices Σ_i for each V_i and then compute

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

where the median is taken with respect to Frobenius norm $||A||_F = \text{trace}(A^\intercal A)$. We take the eigenvectors of $\hat{\Sigma}$, $\{w_1, \ldots, w_6\}$, arranged in order from the largest to the smallest eigenvalue. Then, our robust principal geodesic directions in the tangent space at μ^* are $\{v_1, \ldots, v_6\}$, where v_i is the vector corresponding to w_i, by the vec operator. To form explanatory subspaces, we then exponentiate the span of $\{v_1, \ldots, v_k\}$ at μ^*, for $k = 1, \ldots, 6$.

This procedure is robust, because it ensures that the located center of the data and the located explanatory directions are not unduly affected by 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 κ a scaling parameter. We repeat the experiment of Simulation 1 of section 4.1.1, with each average taken over 1,200 runs.

The results are shown in Table 4, showing again that the median-of-means estimator always improves on the average of the means, and almost always on the overall sample Fréchet mean. The average distance from the truth of the median-of-means for $m = 5, 10, 15$ improves on 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 follows this distribution, then $\text{vec}(\text{Log}_I(X)) \sim \mathcal{N}(\mathbf{0}, \kappa\mathbf{\Sigma})$, with κ a scaling parameter. Σ is diagonal, with diagonal entries varying from 1 to 20 to ensure that population PGA directions exist.

Over 200 runs, we add $0, 5, 10$, and 15 outliers outside a 95% confidence region in $n = 60$ data points, and compute the 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 one,

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

k	$\overline{\rho(\hat{\mu},\mu)}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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	$\overline{\rho(\mu^*,\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$	$\overline{\rho(\hat{m},\mu)}$	$\overline{\overline{\rho(\mu_i,\mu)}}$
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)	

two, and three dimensions for PGA and for RPGA, computed using 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 that of RPGA. Note that RPGA with $m = 1$ groups is the linear approximation of the PGA procedure given in Fletcher and Joshi (2007).

4.3. Hand-shape data in Σ_2^K

We consider the hand-shape data set in Cootes et al. (1995) of 18 hands, with each hand in a planar shape space Σ_2^{72}. A planar shape Σ_2^K consists of objects with K landmarks in $\mathbb{R}^2$ modulo the Euclidean motions rotation, scaling, and translation (Bhattacharya and Bhattacharya (2012); Kendall (1984)). As in Fletcher, Venkatasubramanian and Joshi (2008), we use ellipses as outliers, with each one given as

$$\left\{a\cos\left(\frac{k\pi}{36}\right), b\sin\left(\frac{k\pi}{36}\right); k = 0,\ldots,71\right\},$$

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 three. We then compute and observe the geometric median and the sample mean.

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

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

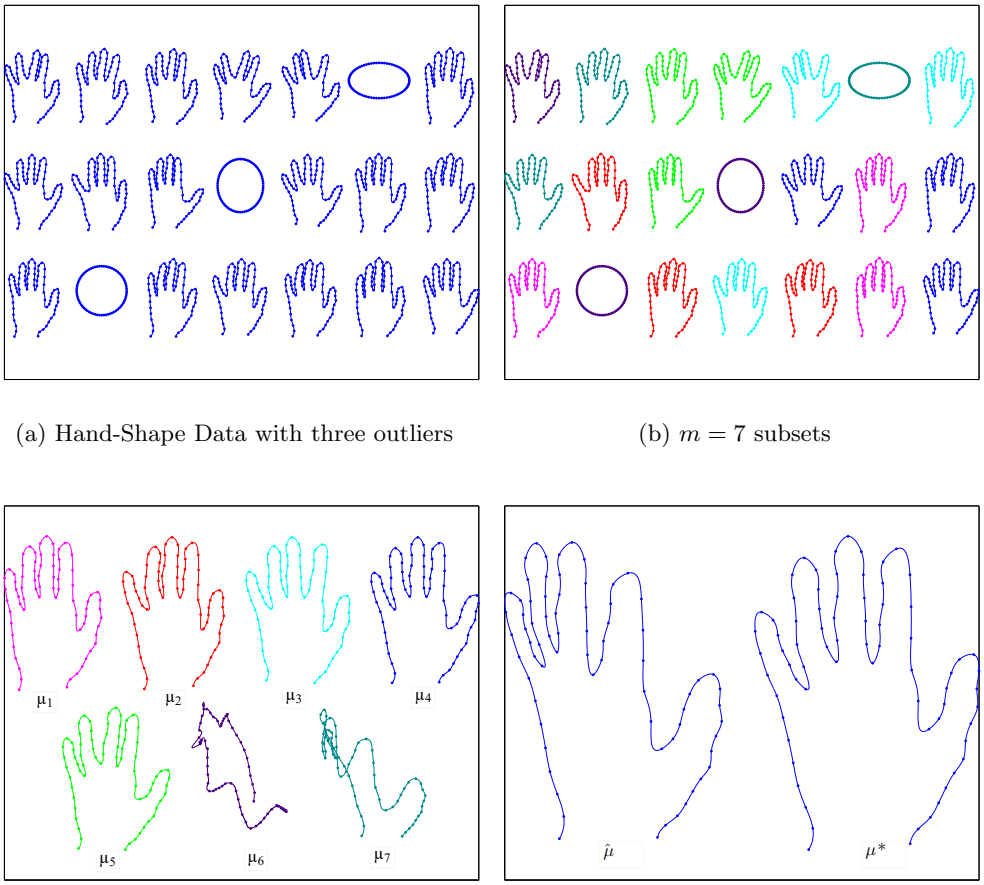
4.3.1. Computation of sample statistics on $\boldsymbol{\Sigma_2^K}$

We identify Σ_2^{72} with S^{69}/S^1, as in (2.4), and compute the intrinsic sample means and medians using direct modifications of the algorithms in Section S2.

In Figure 3 (a), we show $n = 21$ hands with three outliers. In (b), we show seven randomly assigned subsets, indicated by 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, because it retains the shape of a hand similar to the original 18 hands.

5. Discussion

We propose a robust and scalable procedure for general optimization problems on manifolds. Scalability is particularly important for handling the difficult computational issues that arise when estimating sample statistics for manifold data, or when extracting a low-dimensional manifold in high-dimensional data. Note that parallel computation can be implemented trivially from the subsampling procedure.



(a) Hand-Shape Data with three outliers

(b) $m = 7$ subsets

(c) Subset means, μ_i

(d) Sample mean, $\hat{\mu}$, and geometric median, μ^*

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

Lemma 1 provides an important property of geometric medians on manifolds. Then, Theorem 1 shows that the resulting estimator yields provable robustness and tighter concentration bounds about the true parameter of interest. Numerical results from simulated and real data, discussed in Section 4, agree with the robustness and concentration properties of the estimator.

Future research might consider the optimal numbers and sizes of subgroups for estimation, as discussed in remark 2. In Theorem 1, for a given ϵ, additional groups provide a larger m, but also a larger η in the bound provided by (3.4). This is reflected in Examples 1 and 2 in (3.5) and (3.6). Thus, determining the optimal m requires that we also consider the number of outliers and the amount of contamination in the data. In addition, we need to consider the challenging computations for large data sets on manifolds and the advantages of partitioning the data. Furthermore, the second step of the RPGA procedure in 4.2.2 could

partition the data in the manifold, rather than their Riemannian logs, in the tangent space at μ^*. Computing the RPGA, as formulated in 4.2.2, requires only that we compute the median-of-means μ^*, and then use the linear operation of computing the sample covariance matrices of the Riemannian logs of the data in the tangent space at μ^*. Robust estimation on manifolds in other contexts, such as manifold regression (Aswani, Bickel and Tomlin (2011)) might also be considered. As when estimating the mean, additional complications arise in the more general context of a manifold.

Supplementary Material

This Supplementary Material contains the proofs for Theorem 1 and propositions 1, 2 and 3. Also, the algorithms used for computing the various sample statistics in Section 4 are given.

Acknowledgments

We thank the editor, the associate editor and the reviewers for their valuable comments. The contribution of Lizhen Lin was funded by NSF grant CAREER 1654579 and DMS 2113642. The contribution of BS and LL is funded by NSF IIS 1663870, DMS CAREER 1654579 and a DARPA grant N66001-17-1-4041.

Appendix: Proof of Lemma 1

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

$$dL_{J(p^*)}(v) = \lim_{t\to 0^+} \frac{L\left(\gamma(t)\right) - L\left(\gamma(0)\right)}{t} = \lim_{t\to 0^+} \frac{L\left(\gamma(t)\right) - L(J(p^*))}{t} \geq 0 \qquad \text{(A.1)}$$

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

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

where $\mathcal{P}$ is the projection of $\mathbb{R}^D$ onto $\tilde{\mathcal{M}}$, that is, $\mathcal{P}(x) = \operatorname{argmin}_{y\in\tilde{\mathcal{M}}}\rho(y, x)$. 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

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

which will be needed in determining the constant C_α. One can see that

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

Let

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

Then

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

One has

$$\lim_{t\to 0^+} \left(\|\gamma(t) - J(p_j)\| + \|\gamma(0) - J(p_j)\|\right) = 2\|\gamma(0) - J(p_j)\|. \tag{A.2}$$

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,

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

Thus, by (A.2) and the above equation, if $J(p_j) \neq J(p^*)$, one has

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

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

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

Therefore,

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

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

$$\frac{dL_{p^*}(v)}{\|\gamma'(0)\|} = \sum_{j=1}^{m} \lim_{t\to 0^+} \frac{A_j}{\|\gamma'(0)\|} = \sum_{j:p_j\neq p^*} \frac{\langle \gamma'(0), J(p^*) - J(p_j)\rangle}{\|\gamma'(0)\|\|J(p^*) - J(p_j)\|} + \sum_{j=1}^{m} I(p_j = p^*). \tag{A.3}$$

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 α portion of elements of $p_1, \ldots, p_m$ which are at least ϵ distance away from ω, that is, without loss of generality,

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

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

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

and so

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

Notice that

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

Therefore,

$$\begin{aligned}
&\cos\Big(\angle\Big(\mathcal{J}\big(J(\omega)\big) - J(p^*), J(p_j) - J(p^*)\Big)\Big) \\
&\ge \cos\Big(\psi + \angle\big(J(\omega) - J(p^*), J(p_j) - J(p^*)\big)\Big)
\end{aligned}$$

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

We have

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

Then for any $\alpha \in (0, \cot\psi \tan(\psi/2))$ from (A.3)

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

when

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

which is a contradiction with (A.1).

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

$$dL_{p^*}(v) = \lim_{t\to 0} \frac{L(\gamma(t)) - L(\gamma(0))}{t} = \lim_{t\to 0} \frac{L(\gamma(t)) - L(p^*)}{t} \geq 0. \tag{A.4}$$

Denote

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

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

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

Then

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

We see that

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

On the other hand,

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

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

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

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

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

Therefore,

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

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

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

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 $C_\alpha \geq K(1-\alpha)\sqrt{1/(1-2\alpha)}$, which leads to a contradiction with (A.4).

References

Alexander, A. L., Lee, J. E., Lazar, M. and Field, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics* **4**, 316–329.

Aswani, A., Bickel, P. and Tomlin, C. (2011). Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics* **39**, 48–81.

Bhattacharya, A. and Bhattacharya, R. (2012). *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*. Cambridge University Press.

Bhattacharya, R. and Lin, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. In *Proceedings of the American Mathematical Society* **145**, 413–428.

Boumal, N. and Absil, P.-A. (2015). Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications* **475**, 200–239.

Cootes, T., Taylor, C., Cooper, D. and Graham, J. (1995). Active shape models–their training and application. *Computer Vision and Image Understanding* **61**, 38–59.

Dai, W., Kerman, E. and Milenkovic, O. (2012). A geometric approach to low-rank matrix completion. *IEEE Transactions on Information Theory* **58**, 237–247.

Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.* **87**, 250–262.

Fletcher, P. T., Venkatasubramanian, S. and Joshi, S. C. (2008). 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. Anchorage, Alaska.

Fréchet, M. (1948). Lés élements aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré* **10**, 215–310.

Jung, S. (2010). Random number generatrion form von mises-fisher distribution. Technical report. University of Pittsburgh, Pittsburgh.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics* **30**, 509–541.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* **16**, 81–121.

Kolaczyk, E. D., Lin, L., Rosenberg, S., Walters, J. and Xu, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics* **48**, 514–538.

Lazar, D. and Lin, L. (2017). Scale and curvature effects in principal geodesic analysis. *Journal of Multivariate Analysis* **153**, 64–82.

Lin, L., Rao, V. and Dunson, D. B. (2017). Bayesian nonparametric inference on the Stiefel manifold. *Statistics Sinica* **27**, 535–553.

Lin, L., Thomas, B. S., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* **112**, 1261–1273.

Lohit, S. and Turaga, P. K. (2017). Learning invariant Riemannian geometric representations using deep nets. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 1329–1338.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli* **21**, 2308–2335.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning* (Edited by E. P. Xing and T. Jebara), 1656–1664. PMLR.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2017). Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research* **18**, 1–40.

Najfeld, I. and Havel, T. F. (1995). Derivatives of the matrix exponential and their computation. *Advances in Applied Mathematics* **16**, 321–375.

Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, Inc.

Saparbayeva, B., Zhang, M. M. and Lin, L. (2018). Communication efficient parallel algorithms for optimization on manifolds. In *NeurIPS18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3578–3588.

Sommer, S., Lauze, F. and Nielsen, M. (2010). The differential of the exponential map, Jacobi fields and exact principal geodesic analysis. *CoRR, abs/1008.1902*.

Lizhen Lin

Department of Mathematics, The University of Maryland, College Park, MD 20742, USA.

E-mail: lizhen01@umd.edu

Drew Lazar

Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA.

E-mail: dmlazar@bsu.edu

Bayan Saparbayeva

University of Rochester, Rochester, NY 14627, USA.

E-mail: Bayan_Saparbayeva@urmc.rochester.edu

David Dunson

Department of Statistical Science, Duke University, Durham, NC 27708, USA.

E-mail: dunson@duke.edu

(Received May 2022; accepted October 2022)